Rijks*universiteit* Groningen

Genomics
in *Bacillus subtilis*

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijks*universiteit* Groningen
op gezag van de
Rector Magnificus, Dr. D. F. J. Bosscher,
in het openbaar te verdedigen op
maandag 10 mei 1999
om 16.00 uur

door

Michiel Andries Noback

geboren op 20 januari 1969
te Sialkot (Pakistan)

Promotor:	Prof. Dr. G. Venema
Co-promotor:	Dr. S. Bron
Referent:	Dr. J. Kok

VOORWOORD

A small step for mankind, but a giant leap for M.A.N.!

Vijfeneenhalf bewogen promotie-jaren zijn nu (bijna) afgesloten. Dit moment heeft af en toe héél erg ver weg geleken en toch, onverwachts bijna, is het er. Wat nu nog rest is afronden, inpakken en wegwezen. Verrekijker, slaapzak en pen en papier in de knapzak en dan de zonsopgang tegemoet.

Tijd dus om al de mensen die op één of andere manier iets aan de totstandkoming van dit proefschrift hebben bijgedragen te bedanken. Ten eerste natuurlijk Gerard Venema, die mij de mogelijkheid heeft gegeven in de vakgroep moleculaire genetica aan mijn promotie-onderzoek te werken en mij tijdens zeer pittige edoch vruchtbare discussies heeft behoed voor de nodige wetenschappelijke uitglijders. Ook belangrijk was mijn begeleider Sierd-altijddrukdrukdruk-Bron, voor het nakijken van mijn manuscripten en het bieden van een luisterend oor wanneer dat nodig was. Peter Terpstra, een geval apart, schijnbaar bedrevener in de communicatie met computers dan met mensen. Zonder jou zou ik dit voorwoord niet zitten schrijven; je bijdrage in het geheel is onbetaalbaar, hetgeen dan ook wordt gereflecteerd in de vier publikaties waar we beiden als auteur boven staan. Bijna zou ik hier Siger vergeten te noemen. Dit komt echter niet doordat jouw bijdrage zo gering was (integendeel), maar omdat je altijd zo verdomd rustig en stabiel bent, en daardoor niet echt direct opvallend aanwezig.

Dankzij de internationale aard van het project heb ik ook veel nuttige en plezierige contacten opgebouwd met wetenschappers in het buitenland. Ten eerste met Frank Kunst in Parijs, de coördinator van het sequentie project. In Engeland met Colin Harwood en Noel Carter, die aan het *gtaC* hoofdstuk hebben bijgedragen. In Japan met Kazuo Kobayashi en Noatake Ogasawara, die aan het *hit* hoofdstuk hebben meegewerkt. Met Antoine Danchin, ook in Parijs, die erg nuttig commentaar heeft gegeven voor het aminozuur-analyse hoofdstuk. Als laatste Peter Setlow en Irina Bagyan in Amerika, met wie het *yhcN* hoofdstuk tot stand kwam. Ook binnen de vakgroep, met Harold Tjalsma, is het nog tot gezamenlijke publikaties gekomen dankzij het opduiken van nóg een signaalpeptidase. En natuurlijk met Rense Kiewiet, die een belangrijke bijdrage voor de biochemische karakterisatie van het *hit* eiwit leverde, maar vooral een maatje was bij het koffie + peukje + praatje-gebeuren. Bedankt voor alles Rense.

Studenten zijn er ook geweest: Robèr-immer mit der ruhe-Kemperman, Matthijs-kan ik niet alléén met computers werken en jou het pipetteerwerk laten doen?-Kooi, Romke-grote praat klein hartje-Ribbels en Aalke Drijfholt, die ook inzag dat bij 30+°C de Hoornse Plas de beste plek voor een werkbespreking is. Ik hoop dat jullie minstens even veel van mij geleerd hebben als ik van jullie.

De mensen van mijn labzaal. Papa Steven, verslaafd aan discussie zoals ik aan de peuken, praatpaal voor mijn wetenschappelijke en persoonlijke sores (het leven van een Jelojelua gaat bepaald niet over rozen), heel erg bedankt voor alles. Verder Rob, Caroline,

Jos, Emmo en alle studenten die de afgelopen jaren een tijd bij ons verbleven. De computer is nu vrij!

De 'infrastructuur' van de vakgroep was natuurlijk perfect geregeld door Arie (regelaar), Mozes (media), Henk (foto's & dia's), Peter (glaswerk), Peter en Cees (isotopenlab), Anne en Maarten (computerproblemen) en Emma (receptie).

Verder wil ik al de mensen op het lab bedanken die op één of andere manier mijn verblijf in Haren gezelliger, succesvoller, sportiever, inspirerender, of welke -er dan ook, hebben gemaakt. Bert-Jan, Erwin & Douwe voor de gezellige biljardavondjes; Anne (& Aldert) voor de fantastische zeilvakantie op de Noordzee; Leendert: mens sana in corporo sauna; en verder Arjen, Jetta, Albert, Oene, Bertus, JanWillem, Arno, Aske, Maarten, Jan-Maarten, Jan (3×), Germaine, Anton, Kees, Igor, Harma en verder iedereen die ik vergeten ben.

Een paar mensen van buiten het lab wil ik hier nog met name noemen. Romke, zonder de OZT, de TTB, of onze fantastische vogeluitstapjes naar de Lauwersmeer (via Post Gaarkeuken natuurlijk), Schier (Transectje doen, Captain Iglo) en Vlieland (Soooooo!, hamburgers bakken om half zeven 's ochtends) zou het ongetwijfeld allemaal veel minder plezierig zijn geweest dit laatste jaar. Bedankt voor je vriendschap; ik hoop dat die blijft! Inge, ondanks dat het allemaal anders liep dan gedacht/verwacht, bedankt voor de tijd die we samen hadden.

Als laatste wil ik mijn familie (pap, mam, Roel, Miranda en Alex) heel erg bedanken voor de steun, het begrip en de liefde die ik van jullie kreeg tijdens deze toch vaak moeilijke periode van mijn leven. Gelukkig waren jullie er om in mij te geloven als ik dat zelf niet meer deed.

De Arend,
Gedragen door onzichtbare macht
hoog boven ons, laag op aard'
zijn wij het waarom hij lacht
zijn schallende roep
symbool van macht.
Eenzaam wezen, scherpe ogen
het doorgronden van een schijnbaar simpele geest
kan iemand op dié kennis bogen?

De prooi, overlevingskans
waar denkt het aan als
hard, diep als een lans
de klauw stopt
zijn levensdans
en het verwordt tot iets, warm en mals.

Is er berusting in het weten
of ultieme angst in het beseffen
een ander leeft
door mij te eten.

Contents

ISBN: 90-367-1071-5

Printed by: PrintPartners Ipskamp, Enschede

The studies described in this thesis were carried out at the Department of Genetics of the University of Groningen, The Netherlands. The research was financially supported by grants (BIO2-CT93-0272 & BIO4-CT96-0655) from the European Union.

CHAPTER I

Introduction:
Structural and functional genomics

I.1. General introduction

A famous researcher once said: “Give me your DNA sequence and I know you”. How wrong he was! Of course, the genomic DNA sequence of an organism specifies the characteristics of its owner, but knowing this string of letters by no means tells you the exact nature of the corresponding organism. In reality, a long and difficult road separates the determination of the last base of a genome and the understanding of how the organism in question evolved, grows, propagates, interacts with its environment and, eventually, dies. This understanding will be the Holy Grail of generations of geneticists to come.

The term genomics, introduced in 1986 by T. H. Roderick to describe the study of complete genomes, refers to the kind of research that is used to get answers to the above questions. However diverse the interpretation of this term in the scientific community may be, the following definitions more or less cover the subject. Structural genomics is the scientific discipline of mapping, sequencing and analysing genomes, while functional genomics refers to analysis of genome function. The structural genomics phase has a clear end-point with the completion, annotation (and publication) of a genome sequence. The fundamental strategy of functional genomics is expanding the scope of research from studying single genes or proteins to studying all genes or proteins simultaneously in a systematic fashion, in order to obtain a panoramic view of the organisms’ genetic potentials (Hieter & Boguski, 1997).

Like the term, the field of research now referred to as genomics is very young. In fact, until the completion of the first genome sequence in 1995, that of the bacterium *Haemophilus influenzae*, performing genomics research was not really feasible. However, nineteen genomes are already known today, januari 1999, including the genomes of *Bacillus subtilis* and *Escherichia coli*, the model organisms for Gram-positive and Gram-negative bacteria, respectively. The genomes of the model organisms for genetic research, the nematode *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, the human, and the mouse genome, will have been determined within the next decade. This awesome avalanche of information yields, together with technological advances of the past few years, such as DNA chip technology, unprecedented opportunities for the scientific community

to study genetic organisations and regulation processes at a genomic scale. Moreover, the -comparative- study of complete genomes has, and will undoubtedly continue to do so in the future, uncovered biological phenomena that were not even thought of in the past. The genome sequence of the Gram-positive bacterium *Bacillus subtilis*, the subject of this thesis, is a very valuable asset in this study, since it has already been the subject of intense genetic research for several decades.

One of the important lessons geneticists have learned by now is that a chromosome, or genome, is not merely a carrier of a physically bundled collection of genes of an organism. Instead, it reflects the superimposition of a myriad of biological phenomena that have evolved during four billion years or so of the struggle of life on earth.

I.2. Genome sequencing: structural genomics

Completed genomes

The past few years have brought a new revolution in genetics, since it caused a reversion of the way genetic research is being done. Rather than attempting to isolate a gene on the basis of a phenotype, investigators now take the sequence of a gene responsible for the sought-for function from one organism and search its counterpart in the genome of the organism of interest, and then proceed to confirm this function. The year 1995 represents an important landmark of this revolution with the first publication of the DNA sequence of an entire genome of a free-living organism. This concerned the 1,830,137 bps chromosome of the bacterium *Haemophilus influenzae* (Fleischmann *et al.*, 1995). Since then, the number of entirely sequenced genomes has expanded with astonishing speed. The second genome, from the bacterium *Mycoplasma genitalium*, was published 3 months later in that same year. This genome represents the smallest known genome of a free-living organism, consisting of only 580.070 bps and comprising no more than 470 predicted genes (Fraser *et al.*, 1995). The next year brought the completion of the genomes of the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996), the cyanobacterium *Synechocystis* sp. strain PCC6803 (Kaneko *et al.*, 1996), the pathogenic bacterium *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), and -the first eukaryote- *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996; Mewes *et al.*, 1997). In December 1998, the first complete genome sequence of a multicellular eukaryote, that of the 97 Mb genome of *Caenorhabditis elegans*, was published. Table I.1 summarises some basic characteristics of the genomes, the sequences of which have been determined so far. To date, nineteen genome sequences are completely known, and many more are in the process of being determined. Sequencing efforts are presently primarily directed to pathogenic organisms, such as *Enterococcus faecalis*, *Legionella pneumophila*, *Mycobacterium leprae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Vibrio cholerae*. However, most classical subjects of biological study are also being determined, such as the human genome and those of *A. thaliana* and *D. melanogaster*.

Table I.1. Overview of completely sequenced genomes (in order of completion)

Organism Strain ***(% G/C)***	***size (kbp)***	***Number of ORFs (and RNA species)***	***c.d.*** [$]	***% unknown (% w.o. DB match; % sim. to hypoth.)***	***Reference***
Haemophilus influenzae Rd KW20 ***(38%)***	1,830	1,743 (72)	85	42 (22 ; 20)	(Fleischmann *et al.*, 1995)
Mycoplasma genitalium G-37 ***(32%)***	580	470 (36)	88	32 (20 ; 12)	(Fraser *et al.*, 1995)
Methanococcus jannaschii [1] DSM 2661 ***(31%)***	1,664	1,738 (43)	??	38 (22 ; 16)	(Bult *et al.*, 1996)
Synechocystis *sp.* PCC6803	3,573	3,168 (??)	??	56 (45 ; 11)	(Kaneko *et al.*, 1996)
Mycoplasma pneumoniae M129 ***(40%)***	816	677 (39)	89	26 (16 ; 10)	(Himmelreich *et al.*, 1996)
Saccharomyces cerevisiae [2] S288C	12,068	5,885 (455)	70	29 (16 ; 13)	(Goffeau *et al.*, 1996), (Mewes *et al.*, 1997)
Helicobacter pylori 26695 ***(39%)***	1,668	1,590 (44)	91	(31; ??)	(Tomb *et al.*, 1997)
Escherichia coli K-12 ***(51%)***	4,639	4,288 (107)	89	38 (32 ; 6)	(Blattner *et al.*, 1997)
Methanobacterium thermoautotrophicum ΔH ***(50%)***	1,751	1,855 (47)	92	54 (27 ; 28)	(Smith *et al.*, 1997)
Bacillus subtilis 168 ***(44%)***	4,215	4,221 (121)	87	42 (26 ; 16)	(Kunst *et al.*, 1997)
Archaeoglobus fulgidus VC-16, DSM4304 ***(49%)***	2,178	2,436 (56)	93	53 (26 ; 27)	(Klenk *et al.*, 1997)
Borrelia burgdorferii [3] B31 ***(29%)***	911	853 (41)	94	41 (29 ; 12)	(Fraser *et al.*, 1997)
Aquifex aeolicus [4] VF5 ***(43%)***	1,551	1,512 (51)	93	44 (27 ; 17)	(Deckert *et al.*, 1998)
Mycobacterium tuberculosis [5] H37Rv ***(66%)***	4,412	3,924 (50)	91	60 (16 ; 44)	(Cole *et al.*, 1998)
Treponema pallidum *pallidum* ***(53%)***	1,138	1,041 (50)	93	45 (28 ; 17)	(Fraser *et al.*, 1998)
Pyrococcus horikoshii OT3 ***(42%)***	1,739	2,061 (50)	91	80 (58 ; 22)	(Kawarabayasi *et al.*, 1998)
Chlamydia trachomatis (41%) [6] serovar D (D/UW-3/Cx)	1,043	894 (??)	??	32 (28 ; 4)	(Stephens *et al.*, 1998)
Rickettsia prowazekii Madrid E ***(29%)***	1,112	834 (37)	76	38 (25 ; 13)	(Andersson *et al.*, 1998)
Caenorhabditis elegans [7]	97,000	19,099 (>1000)	27	?? (58 ; ??)	(*C. elegans* sequencing consortium (see genome.wustl.edu/gsc/C_elegans), 1998)

[$] c.d.: *coding density*: the percentage of DNA that is actually coding for proteins and RNA species.

??: Not presented in corresponding publication.

[1] *M. jannaschii* contains two extrachromosomal elements (ECEs) of 58 kb (28% G/C) and 16 kb (29% G/C), respectively. These data are the sum of the three genomic elements.

[2] The yeast genome consists of 16 chromosomes and only 12.068 kb of the genome, totalling 13.389 kb, was sequenced. Not completely sequenced are Ty-elements, rDNA repeats, mtDNA and some highly repeated genes (e.g. CUP1 & ENA2). Information on similarities was obtained from: http://www.mips.biochem.mpg.de/yeast/

[3] *B. burgdorferi* contains one linear chromosome of 910,725 bps and at least 17 linear and circular plasmids totalling over 533,000 bps. Presented information covers only the chromosome. The

plasmids, with a coding density of 71%, encode at least 430 polypeptides of which 58% has no d.b. match and 26% match only hypothetical proteins.

[4] *A. aeolicus* contains one ECE of 39,5 kb with a G/C content of 36,4% and a coding density of 54%.

[5] The 44% indicated under "similar to hypotheticals" also includes weak similarities.

[6] *C. trachomatis* contains one ECE of 7,5 kb.

[7] The genome size is an approximation, since some gaps remain in the sequence. 16,260 *C. elegans* genes have been reviewed for these data. On average, there are 5 introns per gene, and 27% of the DNA encodes exons.

Information on genome projects can be obtained at:
MAGPIE: http://www.mcs.anl.gov/home/gaasterl/genomes.html
TIGR organization (The Institute for Genomic Research): http://www.tigr.org.
Genome sequencing centre (C. elegans): http://genome.wustl.edu/gsc/C_elegans/

The *A. thaliana* genome is scheduled to be finished in 2000 (Meinke *et al.*, 1998) and the human genome, comprising three billion basepairs, is scheduled to be entirely sequenced by the end of the year 2003 (Collins *et al.*, 1998). Since the *D. melanogaster* genome will be finished by 2002, and the mouse genome project being started in the near future, the genomes of all important model organisms used in genetic research will be determined in the course of the next decade.

Viral (& phage), mitochondrial, and plastid genomes will not be discussed here, although over 200 viral, 20 mitochondrial, and 11 plastid genomes have now been determined and published.

Representatives from all three major kingdoms, eukarya, bacteria, and archaea, have now completely been sequenced, presenting unprecedented opportunities for comparative genome analyses. From Table I.1 it is evident that bacterial genomes all have a gene-coding density of about one gene per kilobase of DNA and, except for *Rickettsia prowazekii*, a coding density of about 85-90%. Another general feature of the sequenced microbial genomes is the finding that about 20-30% of their genes are unique, *i.e.* the databases do not contain related genes from other organisms (orthologs). The genome of *P. horikoshii* is an exception to this rule, with almost 60% of its putative protein sequences being unique. This is probably a result of the fact that this species is a hyper-thermophilic archaebacterium, growing optimally at temperatures of nearly 100°C (Kawarabayasi *et al.*, 1998). The other thermophilic archaebacteria, the genomes of which have been determined, *M. jannaschii*, *M. thermoautotrophicum*, and *A. fulgidus*, have much lower optimal growth temperatures, of 85°C, 65°C, and 83°C, respectively. Apparently, with growth temperatures approaching 100°C, this poses unique constraints on protein architecture. Since the fraction of protein sequences without database matches in a genome have not significantly decreased with each new sequenced microorganism, this is likely to indicate that about one quarter of these organisms' genes are probably specifying the unique character traits typical of that organism.

The genomes of *H. influenzae*, *M. genitalium*, *M. tuberculosis*, *M. jannaschii*, *C. trachomatis*, *R. prowazekii*, *T. pallidum*, *A. aeolicus*, *H. pylori*, *M. thermoautotrophicum*, *A. fulgidus*, and *B. burgdorferii* have been determined by a strategy called whole-genome random

sequencing (see below). The *B. subtilis* genome, however, like the genomes of *P. horikoshii*, *Synechocystis*, *S. cerevisiae*, *E. coli*, has been determined by a directed approach. All research groups involved in the *B. subtilis* genome-sequencing project were assigned a particular region from the genetic map of the chromosome (Anagnostopoulos *et al.*, 1993). Cloning of the assigned region was carried out by various positional cloning methods, including plasmid walking, marker rescue, lambda bank screening, and (inversed) long-range PCR (Cheng *et al.*, 1994).

Whole-genome random sequencing

The strategy of choice for the sequencing of an entire genome is presently whole-genome random sequencing. The basic principle behind this method is a modification of the Janus strategy (Burland *et al.*, 1993) and involves the construction of two independent chromosomal DNA banks: a plasmid bank with relatively short inserts and a bank in a phage λ derivative with large inserts. Subsequently, high-throughput DNA sequencing of clones from primarily the first bank is performed, and this is supplemented by sequences obtained from the second, λ bank. The plasmid bank generates the main body of the sequence, while the λ bank is used primarily for controlling the physical integrity of the sequence and as a source of linking clones for the determination of physical gaps in the sequence after assembly of sequences obtained from the plasmid bank. This strategy is particularly powerful when the organism to be sequenced has not been extensively studied, and hence no genetic and physical maps of the chromosome are available at forehand. Most known genomes have been determined by this strategy (Fleischmann *et al.*, 1995).

The essence of this approach consists of several, partially overlapping phases, summarised in Table I.2. In the first phase, BAL 31 nuclease-generated random genomic libraries are constructed in a high-copy number plasmid (e.g. pUC18) for *E. coli*, and in a phage lambda derivative (*e.g.* λ GEM-12 or λ DASH II). The plasmid shotgun bank contains relatively small DNA fragments in the 1,6 to 2,0 kb range, and the lambda genomic library contains fragments in the 15-20 kb range. These are verified for their random coverage of the genome.

The second phase consists of high-throughput DNA sequencing and assembly, primarily of clones from the plasmid library, and complemented by sequences from the lambda library. To estimate the amount of DNA that needs to be sequenced to yield the desired coverage of the genome, the following equation for the Poisson distribution can be used: $P_O = e^{-m}$, where m is the sequence coverage, and P_O is the probability that a base will not be sequenced. Thus, with *B. subtilis* as an example (genome size 4.2 Mb), after sequencing 4.2 Mb from random clones, $P_O = e^{-1} = 0.37$, which means that 37 percent of the genome is expected to be unsequenced. If L is the genome length and n is the number of random sequences generated, the total gap length is Le^{-m}, and the average gap size is L/n. So, again using *B. subtilis* as example, fivefold coverage (21,000 clones sequenced from both ends with an average of 500

bases), would yield a total gap length of 28,299 bases and an average gap size of about 100 bases.

In the third phase, contigs are ordered and the remaining sequence gaps are closed. This is done by primer walking, primarily from linking clones in the lambda bank.

Table I.2. Stages of the whole-genome sequencing strategy (adapted from Fleischmann *et al.*, 1995)

Stage	Description
Random small insert and large insert library construction	Shear genomic DNA randomly to ~2 kb and 15 to 20 kb fragments, respectively
Library plating	Verify random nature of library and maximise random selection of small insert and large insert clones for template production
High-throughput DNA sequencing	Sequence sufficient number of fragments from both ends for about a 6× coverage of the entire sequence
Assembly	Assemble random sequence fragments and identify repeat regions
Gap closure	
Physical gaps	Order all contigs (fingerprints, peptide links, λ clones, PCR) and identify templates for closure
Sequence gaps	Complete the genome sequence by primer walking
Editing	Inspect the sequence visually and resolve sequence ambiguities, including frameshifts
Annotation	Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions)

Developments in sequencing techniques

The advent of genome sequencing projects has been entirely dependent on recent developments in sequencing techniques; a decade ago such efforts would not have been practically feasible undertakings. Therefore, a short overview will be given on developments from the Maxam & Gilbert (1977) chemical sequencing method to present-day, fully automated systems based on the dideoxy chain-termination –or enzymatic- method of Sanger *et al* (1977). Both methods produce nested sets of (radioactively) labelled polynucleotides, from 1 to 500 bases long, that begin at a fixed point and terminate at points that depend on the location of a particular base in the original DNA strand. The polynucleotides are then separated by polyacrylamide gel electrophoresis (PAGE), and the order of nucleotides in the original DNA can be read directly from an autoradiograph, or a fluorogram in the case of fluorescent labelling (Griffin & Griffin, 1993).

Cycle sequencing is an adaptation of the Sanger dideoxy method of sequencing. It has the advantages that reactions are simpler to set up, less template is required, and the quality and purity of the template are not as critical as in the standard procedure. In this method, using polymerase chain reaction (PCR) technology, a single primer is used to amplify the region to be sequenced in a linear manner using *Taq* DNA polymerase in the presence of deoxy-nucleotide triphosphates (dNTP's) and a dideoxy-nucleotide triphosphate (ddNTP).

Automated DNA sequencing has become one of the major advancements in modern biotechnological research. It is based on the Sanger chain termination method of sequencing, but uses fluorescent instead of radioactive labelling techniques. The label is attached either to the sequencing primer, the nucleotides, or the dideoxy nucleotides. During electrophoresis of the DNA fragments in a PAA gel, the fluorescent label is excited by a laser beam and the emitted fluorescence is collected by detectors. The fluorescence signal that is generated is analysed by a computer. Two approaches of this basic concept are presently employed. In the first, one type of label is used for all four sequencing reactions (A, C, G and T), and the reactions are run in separate lanes of the sequencing gel. In the other approach, four different types of label, each having different fluorescence wavelengths, are used in the sequencing reactions and the products of the four reactions are run in one lane on a sequencing gel. Signal detection is then performed at the four different fluorescence wavelengths of the labels.

Today, sequencing systems exist that are fully automated through almost all stages of the process: from toothpicking colonies off a plate, culturing cells, extracting DNA, and sequencing of the template, to the computer-assisted assembly of the raw sequence data.

Although the concept itself is not new (Drmanac *et al.*, 1993), sequencing by hybridization (SBH) to an array of oligonucleotides has only recently become feasible with the advent of the DNA chip technology (Drmanac *et al.*, 1998). This methodology is discussed in the paragraph dedicated to applications of the DNA chip technology (see below).

Finding structural features in a DNA sequence

Once a DNA sequence has been completed, the annotation phase begins. The aim of this phase, which consists of a logical order of analyses, is to identify as many as possible primary structural features within the DNA that has been sequenced. This includes identification of open reading frames (ORFs) and verification of codon usage, identification of the start (start codon + ribosomal binding site) and stop sites of ORFs, and analysis of possible terminator structures and promoters. All the analyses discussed below are performed with the aid of computer programs designed for that purpose.

First, ORF searches are performed, and the ORFs that are found are subjected to several filtering procedures that are meant to identify the ones that are likely to constitute protein coding regions. One should keep in mind that all these filtering procedures are imperfect, and also tend to reject several ORFs that are in fact real genes (Bains, 1992). ORFs are first selected on the basis of their length. Usually, a cut-off value of 50 to 100 codons is employed, implying that smaller genes will be missed. Genes encoding RNA species (rRNA, tRNA) are identified by sequence similarity.

The second step in gene annotation is the identification of translational start and stop signals. A typical bacterial translational start site consists of a Shine-Dalgarno (SD) sequence, also called Ribosomal Binding Site (RBS), followed within 4-10 basepairs by one of the start codons ATG, TTG, or GTG. The RBS is recognised by the 16S ribosomal RNA, and should be (partially) complementary to its 3’ end. In *B. subtilis*, the RBS should preferably contain

part of the sequence 5'-AGAAAGGAGGTGATC-3'. Although ATG (78 %), and to a lesser extend TTG (13 %) and GTG (9 %) are the main start codons in *B. subtilis*, ATT and CTG have also been identified as the start codon in a small number of genes (Kunst *et al.*, 1997). An ORF ends with any of the three stop codons TAA, TGA, or TAG.

Subsequently, the codon usage of the ORFs is compared with the average codon usage of the organism from which the DNA sequence was obtained. In *B. subtilis*, genes can be separated into three classes according to codon usage (Kunst *et al.*, 1997). Class I comprises the majority of the *B. subtilis* genes (82%), including the genes for sporulation. Class II includes genes that are highly expressed during exponential growth (4.6 %), such as those encoding the transcription and translation machineries, stress proteins, and core intermediary metabolism. Class III genes (13%) are mainly of unknown function and these genes are enriched in A + T residues. They are mostly located in or associated with (remnants of) bacteriophages, transposons, or functions related to the cell envelope. Codon usage tables for many organisms are obtainable via the www (at URL: http://www.dna.affrc.go.jp/~nakamura/CUTG.html).

The fourth step is the identification of structures involved in transcription termination. A (rho-independent) terminator is a sequence element that can form a stem-loop structure; it consists of a short inverted repeat (the stem) separated by a few bases (the loop). Identification of terminators yields insight into the possible transcriptional organisation of genes in operon structures.

A further option is the search for promoters for transcription initiation. This is, however, not an option that is easily performed, since bacterial promoters are too variable in sequence and spacing for unambiguous identification by simple search-strings. This might still be feasible if bacteria would not have possessed multiple RNA polymerase sigma factors with different sequence specificities for their binding. For *E. coli* however, a neural network was trained to recognise σ^A-specific promoter sequences.

Finally, possible functions for ORFs can be deduced by performing homology analyses. When an amino acid sequence displays a high level of similarity to a sequence with known function from another organism, it is very likely that the putative gene from the organism of interest performs the same, or a similar function. When no full-length homology to known protein sequences is observed, identification of functional domains or motifs can be useful in determining at least part of the function of a given gene. Such domains include, for instance, ATPase domains characteristic of ABC transporters, helix-turn-helix domains of DNA-binding domains, signal sequences typical of secreted proteins, transmembrane helices, but also smaller sequence elements, called motifs, like the Walker A ATP-binding motif. However, this kind of analysis already resides in the realm of functional genomics, and therefore will be discussed below.

I.3. Functional genomics

Introduction

The recent explosion in available sequence data has induced the rapid development of a new field of science, now termed genomics. Concomitant with the advent of this research area, the vocabulary of biological terms concerning the classification of proteins has expanded as well. Therefore, a brief overview of this new terminology will be presented here. Like organisms, proteins can be classified on the basis of their relationships. Homologous protein sequences can either be orthologs or paralogs. Orthologs are homologous sequences with a common ancestor, separated by speciation events, which perform the same role in different species. Paralogs are homologous proteins resulting from gene duplications within a species. Normally, orthologs retain the same function in the course of evolution, whereas paralogs may evolve new functions, which may or may not be related to the original one (Tatusov *et al.*, 1997). Thus, paralogs usually perform similar functions, but not necessary the same (Henikoff *et al.*, 1997). When in different organisms the same function is performed by nonorthologous proteins, this is called a nonorthologous displacement (Mushegian & Koonin, 1996). In the context of protein homologies, the term orphan is also used, to describe a gene product that does not show any similarity to protein sequences from other organisms, or shows similarity only to proteins of unknown function (Dujon, 1998). The complete set of protein sequences encoded by a genome is called the proteome, and the complete set of (possible) RNA transcripts from a genome is called the transcriptome.

Composite proteins consisting of multiple modules, the functional building blocks of proteins, are called chimeras. Modules can contain one or multiple motifs. These smallest units in protein classification are recognised as highly conserved similar regions (amino acids) in alignments, and these often represent an enzyme's active site residues. They can be as small as, for example, the zinc-finger DNA binding motif C_2H_2. By virtue of forming an independently folded structure, this motif is the signature of a module. Motifs can either reflect common ancestry or convergence from different origins (e.g. the Walker A ATP-binding motif).

Major challenges in genomics are the systematic analysis of genome function, regulation, and evolution. Several methodologies have been designed to address these goals, separately or in concert, with single genes or genome-wide, and some of these will be reviewed below.

Analysis of (single) genes: mutant construction & phenotype screening

When the genome of an organism has been annotated and the putative genes are known, the real work in a genomic study has just begun. To uncover the function of all (orphan) genes in a genome, a systematic high-throughput approach for mutant construction and phenotype screening is essential. For the genomes of *B. subtilis*, yeast, and *C. elegans*, programs are now underway to generate strain collections in which all genes -especially the orphan genes- are mutated one-by-one in individual mutant strains. Subsequently, the mutants are systematically screened for a phenotype (Dujon, 1998). In this paragraph, only the projects on the functional analysis in *B. subtilis* and yeast will be discussed.

In order to study gene function in *B. subtilis* in a systematic way, a consortium of research groups in the European Union and Japan was set up for this purpose. The goal of this program was to construct mutants of all genes with unknown function, and to assign functions to these genes by systematic and high-throughput phenotype screening. Proteins without any significant similarity to proteins in the public databases, or with similarity to uncharacterised proteins only, are considered unknown in this project. An integrational vector, pMUTin2, was constructed, that can be used for the construction of insertional mutants. Concomitant with the formation of an insertional mutation, a transcriptional fusion of the promoter of the gene of interest to the *lacZ* reporter gene is generated, such that possible downstream genes in operon are placed under the control of the IPTG-inducible P_{spac} promoter (Vagner *et al.*, 1998). Each participant in the consortium is responsible for the construction of an assigned number of mutant strains and the subsequent analysis of these strains with respect to growth and *lacZ* expression in rich and minimal medium. Also, each mutant strain has to be screened for phenotypic effects with respect to a number of characteristics, as listed in Table I.3.

When mutant strains show a phenotype concerning any of these characteristics, these are subsequently analysed in more detail by members of the consortium which are specialised in that particular area of research. All data are entered in a central database called Micado (at http://locus.jouy.inra.fr/cgi-bin/genmic/madbase_home.pl), which for a limited period of time is accessible only to members of the consortium. By the end of 1998, 1035 mutant strains had been deposited in the *B. subtilis* mutant collection.

Table I.3. Areas of phenotype screening in *Bacillus subtilis* (from Gas *et al.*, 1998)

Small and inorganic molecules	*Stress and stationary phase*	*Cell structure and motility*
Carbon metabolism	Stress analysis	Cell envelope
Nitrogen and sulphate	Stationary phase	Motility
Macromolecules	*Cellular processes*	
DNA	Cell cycle	
RNA	Competence	
Protein	Sporulation	
	Germination	

The yeast *S. cerevisiae* is being subjected to a similar systematic functional analysis program by a consortium of 134 research groups called EUROFAN (European Functional Analysis Network). In this program, orphan ORFs are completely deleted and replaced by a kanamycin marker cassette. The disruptants, when viable in the haploid state, are then subjected to a first level analysis that includes growth on three basic laboratory media at three different temperatures, mating, and sporulation (Dujon, 1998).

Genomics at the transcriptome level: DNA chip technology

The DNA chip technology provides one of the most powerful genetic research tools that have been developed in the past few years. The technique can be used for many purposes, among which are: genome-wide parallel gene expression studies, identification of regulons and their regulatory motifs, gene discovery studies, polymorphism screening, mapping of DNA clones, and DNA sequencing by hybridization (Ramsay, 1998; Drmanac *et al.*, 1998). Another method for analysis of the transcriptome is Serial Analysis of Gene Expression (SAGE; Velculescu *et al.*, 1995). In this method, very short cDNA sequences from a complete mRNA population are concatamerisated in a cloning vector and subsequently sequenced, providing insight in which genes are transcribed and how abundantly. This technique will not be further discussed here.

The principle of DNA chip technology is the following. An ordered array of oligonucleotides or DNA fragments, for instance representing all genes from a genome, is synthesised and immobilised on a solid base. Subsequently, this array is exposed to a fluorescently labelled probe, or a set of differently labelled probes, and fluorescence is monitored at each position of the array. Two basic variants of chip manufactory exist today: delivery or synthesis. In the delivery variant, prefabricated oligonucleotides, or PCR-generated DNA fragments, are immobilised in an ordered array on a solid surface. In the synthesis variant, an array of oligonucleotides is synthesised *in situ* through a combination of photolithography and oligonucleotide chemistry (Ramsay, 1998; Marshall & Hodgson, 1998; Schena *et al.*, 1998).

One important application of this technology is genome-wide simultaneous gene expression monitoring. Here, a chip containing oligonucleotides or complementary DNA fragments, each uniquely complementary to a gene from the genome under investigation, is used. The DNA on this chip is then hybridised with total mRNA from a culture grown in a certain medium or under specific conditions. This analysis yields insights in the transcriptional program used under the particular growth condition, and it can also serve as a reference transcriptome. Subsequently, the same experiment can be performed with mRNA isolated from a culture that is subjected to different growth conditions or medium components. When the transcriptional programs from these two experiments are compared, this identifies genes that are specifically induced or repressed under the conditions tested. The sensitivity of the method, as tested with human, yeast, and bacterial cells, is such that between 0.1 to 5 mRNA copies per cell can be detected (Lockhart *et al.*, 1996; Wodicka *et al.*, 1997; de Saizieu *et al.*, 1998). The

DNA chip technology has been used successfully in several instances. In budding yeast it has been applied to identify genes involved in the transcriptional program of sporulation and, concomitantly, regulator-binding consensus sequences upstream of these genes (Chu *et al.*, 1998). Roth and co-workers (1998) have shown in yeast, using the galactose response, heat shock, and mating type as examples, that alignment of upstream sequences of genes that are expressed in concert readily reveals the consensus regulatory motif of the respective regulon. Also in yeast, genes involved in growth in rich and minimal media were identified (Wodicka *et al.*, 1997), as well as the temporal program involved in the metabolic shift from fermentation to respiration (DeRisi *et al.*, 1997). In this latter paper, another application of the method, the identification of target genes of transcriptional regulators, was also presented. In this case, targets of regulators (*TUP1* and *YAP1*) involved in the shift from fermentation to respiration were identified.

To investigate DNA sequence differences (polymorphisms), two approaches can be used. In the first approach, sequencing by hybridization (SBH), an array is used consisting of a complete set of noncomplementary oligonucleotides. With seven-mers, for instance, 8192 oligonucleotides are necessary in the array. Hybridization of an unknown DNA fragment to this array yields its sequence. Using this method, a number of polymorphisms in a 1.1 kb DNA fragment carrying the human the *p53* gene was determined (Drmanac *et al.*, 1998). In the other approach, an array of oligonucleotides designed to match specific sequences is used. Sequence polymorphisms in human mitochondrial DNA were identified successfully with this approach (Chee *et al.*, 1996).

Genomics on the proteome level: 2D gelelectrophoresis and the two-hybrid system

Two methods exist to perform genome analysis at the protein level; these are two-dimensional (2D) gel electrophoresis and the yeast two-hybrid system.

2D gel electrophoresis is a method that separates proteins in two dimensions. In the first dimension, proteins are separated on the basis of their charge, or isoelectric point, while they are separated on the basis of their mass in the second dimension. This is -in theory- a powerful tool for genome analysis at the protein level. It is, for instance, possible to study changes in protein levels in response to changes in growth conditions, or the effects of mutations in regulators. Progress in this field is, however, hindered by two limitations of the technique. First, only a limited number of protein spots is visible on a 2D gel (around 500 for *B. subtilis* and 1500 for yeast). Second, the number of identified protein spots is only slowly accumulating, since identification has to be done through time-consuming methods such as microsequencing and mass spectroscopy. In *B. subtilis*, attempts have been made to construct such a map, or 2D protein index (Schmid *et al.*, 1997). Also, characterisation of protein changes in response to heat shock, salt- and ethanol stress, is well underway (Bernhardt *et al.*, 1997). Proteome analysis through 2D gel electrophoresis is also employed in the yeast functional genomics program (Dujon, 1998).

The two-hybrid system is a genetic method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature, characteristic of many site-specific transcriptional activators, which consist of a DNA-binding domain and a transcriptional activator domain. The DNA-binding domain targets the activator domain to the specific genes that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription to occur. The two-hybrid system is based on the observation that the two domains of the activator, which is the yeast Gal4 protein in most applications, need not be covalently linked and can be brought together by the interaction of any two proteins. It requires that two hybrids are constructed: a DNA-binding domain fused to some protein, X, and a transcription activation domain fused to some protein Y. These two hybrids are expressed in a cell containing a reporter gene with an upstream consensus sequence for the DNA-binding domain. If the proteins X and Y interact, they create a functional activator, the activity of which can be detected through the expression of the reporter gene (Phizicky & Fields, 1995). In yeast, surveys of protein-protein interactions are already in progress, either using pre-mRNA splicing factors as initial bait (Dujon, 1998), or genome-wide, investigating all 18 million pair-wise combinations of the approximately 6,000 predicted proteins of yeast (Hieter & Boguski, 1997).

I.4. *In silico* genomics

The availability of complete genome sequences makes it possible to investigate the evolutionary relationships between genomes and gene families within a genome. Several aspects can be addressed with *in silico* genomics, at different levels of hierarchy: at the DNA level, the protein level, and level of the organism. A number of examples will be presented in the following paragraph.

In silico DNA analysis

When investigating genome features at the DNA level, searches for evidence of non-randomness of (oligo)nucleotides (words) in a DNA sequence can be made. This non-randomness is then analysed in a realistic model of the genome. The rationale for this is that words that are over- or under-represented in a sequence, in contrast to a model, may indicate a phenomenon of positive or negative selection (Rocha *et al.*, 1998). For example, when investigating oligonucleotide frequencies in a genome, a period of three will be observed. This period is known to be related to the codon size of three, and in a realistic model, therefore, this period of three should be incorporated in the analysis. The same applies to frequencies of nucleotides (the G+C content) and all other known sequence elements that result in non-randomness of the nucleotide composition.

Analysis of codon usage in *E. coli* and *B. subtilis* has revealed the existence of three classes of genes, also distinguishable by their biological properties. The majority of the genes fall into Class I, representing genes that are expressed rarely and/or genes expressed at a low level. Class II consists of genes expressed continuously at a high level during exponential

growth. Class III consists of genes corresponding to surface elements of the cell, genes from mobile elements, as well as genes resulting in a high fidelity of DNA replication. It was suggested that class III genes are inherited through horizontal transfer (Médigue *et al.*, 1991; Kunst *et al.*, 1997).

Rocha *et al.* (1998) have observed biased word usage in the genome of *B. subtilis* with words of up to eight letters long. Biases in trinucleotide frequencies were mainly coupled to codon usage. Biased words of size seven are probably related to interaction with RNA or DNA polymerase, and avoidance of palindromic sequences would be the result of avoidance of restriction sites, since restriction/modification systems are easily horizontally transferred. In *E. coli*, codon usage of genes for major structural components of the outer membrane, porins and lipopolysaccharides, indicate that these genes might be obtained through horizontal gene transfer (Guerdoux-Jamet *et al.*, 1997).

The distribution of the methylation motif GATC in the *E. coli* genome, which is known to be also involved in the long-patch mismatch repair system, revealed regions which are abnormally rich in this motif. These GATC clusters were mainly located within genes involved in the shift from anaerobic to aerobic growth. The GATC clusters probably represent some kind of transcriptional regulation process accompanying the shift of the bacterium from the host environment (high temperature, lack of oxygen, high osmolarity) to the external medium (low temperature, presence of oxygen, low osmolarity). A possible mechanism was presented which includes the fact that methylation lowers the T_m of the DNA helix drastically. Also, a strong bias was observed against GATC motif pairs separated by 6 bases. This under-representation could be explained, since CAP binding sites (TGTGATCTAGATCACA) and FNR binding sites contain this double motif (Hénaut *et al.*, 1996).

In silico protein analysis

At the protein level, paralogous and orthologous relationships can be investigated. Although the definitions of orthologs and paralogs are clear, in practice several problems arise. First, what is the level of amino acid similarity that makes two proteins paralogous to each other? Related to this problem is the question whether paralogs should display similarity over the entire length of their sequence. This is especially problematic in the case of chimeras, composite proteins consisting of multiple domains. It is well known that two separate proteins from one organism can be conserved in another bacterium, where they may be translationally fused to form one protein (Henikoff *et al.*, 1997). Although these proteins obviously have a paralogous relationship from a functional point of view, they will not be designated as such in most studies. Also, proteins can be categorised as paralogs on the basis of their amino acid similarity while they are not functional paralogs. Even identical proteins can, conceivably, be functionally non-paralogous because of differences in temporal and/or spatial expression patterns. A last remark should be made in this respect. Protein sequences from different organisms, though similar, may not be orthologs. This is the case when multiple paralogs are present in one organisms' genome as possible candidates for being the ortholog to a protein

from another organism. In general, though this is not a very reliable rule when investigating species that are phylogenetically very divergent, the pair of sequences displaying the highest similarity to each other are considered orthologs (Tatusov *et al.*, 1997). In genomes, the size of gene families is related to genome size (Huynen & Nimwegen, 1998). In a genome-wide analysis of orthologous and paralogous relationships, Tatusov and co-workers (1997) have identified a total of 720 clusters of orthologous groups (COGs), in which 37% of all the genes from the seven analysed genomes could be classified.

A similar analysis, although on the level of protein structure instead of protein sequence, was performed by Gerstein & Hegyi (1998). These researchers observed that all known protein structures could be classified into a very limited number, currently about 350, of folding patterns. Probably, all proteins occurring in nature are composed of no more than around 1000 folds.

Wallin & Von Heijne (1998) have investigated the occurrence of transmembrane segments in fourteen completed genomes. Twenty to thirty percent of all ORFs were predicted to encode membrane proteins, and this number increases linearly with greater genome sizes. The widespread assumption that organisms have a preference for membrane proteins with 6 or 12 transmembrane segments, was invalidated to some extent, since this preference was observed in some genomes (*E. coli*, *B. subtilis*, *A. fulgidus*, *H. influenzae*, *H. pylori*, *M. genitalium* and *Synechocystis*), but not in others (*S. cerevisiae*, *C. elegans*, *H. sapiens*, *M. thermoautotrophicum*, *M. jannaschii*, *C. acetobutylicum*).

In silico analysis of organisms

At the organism level, several aspects can be studied. For instance, Rosa & Labedan, (1998) have analysed the evolutionary relationship between the bacteria *E. coli* and *H. influenzae* and deduced a putative last common ancestor. The approach used here was an exhaustive analysis of homologous proteins encoded by genes in both genomes through comparison of the evolutionary distances between orthologs and paralogs. Significant similarities were observed between 1,345 *H. influenzae* proteins and 3,058 *E. coli* proteins, many of them belonging to families of various sizes. In other studies, the minimal gene complement for self-sufficient cellular life was deduced from the genomes of two representatives of ancient bacterial lineages, the bacterium *M. genitalium* and the Gram-negative bacterium *H. influenzae* (Mushegian & Koonin, 1996; Koonin & Mushegian, 1996). Only 240 orthologous proteins were found to be encoded by both genomes, and these were assumed to be probably essential for cellular function. This set was complemented with nonorthologous displacements, and removing apparent functional redundancies and parasite-specific genes narrowed the resulting set further down. This yielded a minimal gene complement of only 256 genes. These analyses can be facilitated by a visual representation method called differential genome display, which enables the rapid identification of special features of organisms, such as virulence factors (Huynen, 1997).

An important goal in this type of analyses is, of course, to resolve the ancestry of all life. Attempts in this direction have also been made (Mushegian & Koonin, 1996; Koonin & Mushegian, 1996). Since the minimal gene complement does not contain eukaryal or archaeal homologs of the key proteins of bacterial DNA replication, it seems likely that the last common ancestor of the three primary kingdoms had an RNA genome. Furthermore, as archaea have significant deviations in the enzymology of the upstream reactions of glycolysis, the ancestor may have had a metabolism based on trioses and pentoses as energy source through the conversion to glyceraldehyde 3-phosphate.

I.5. Outline of this thesis

The central question that will be addressed in this thesis is what the scientific spin-off is that can obtained from the vast body of information that genome sequencing projects generate. The *B. subtilis* genomic sequence, part of which was determined in our group, is used as an example to illustrate this general theme.

In chapter one, an overview will be given of strategies and methods that are employed in genome sequencing projects. The central question here is: what is the value of this giant investment of time, money, and human recourses in these seemingly boring and certainly tedious projects. An outline will be presented on how *in silico* analyses can be performed on raw sequence data that are no more than a silent string of letters at the start, but notwithstanding this, define the life of its owner. These analyses can be carried out to address several questions. What can be said about the function of a putative gene, solely by analysing the similarity of its deduced amino acid sequence to other proteins and investigating the presence of motifs, localisation signals and other features? What is its relationship to other proteins in the same and other genomes? What are the implications with respect to insights in the evolution of species, protein paralog and ortholog families, and the biological relevance of these?

In chapter two, the cloning, sequence, and annotation of the *B. subtilis* DNA region that was determined in our group, is presented. This is the chromosomal region located between the *prkA* and *addAB* markers, a DNA fragment of 171,812 bps. This region was obtained through various cloning methods, including lambda bank screens, plasmid rescue, and (inversed) long-range PCR. The *in silico* DNA analyses in this chapter are restricted to gene annotation, terminator searches, and searches for homologs in public databases. Because large differences were observed with respect to existing data, a correction of the original physical and genetic maps of the *prkA* to *addAB* region, originally published by Itaya & Tanaka (1991) and Anagnostopoulos *et al.* (1993), will also be presented.

Chapter three deals with the results of further analyses on the protein sequences encoded by genes in the *prkA* to *addAB* region. These include: paralog frequency analysis, paralog positional analysis, analysis of compartmentalisation signals (membrane-spanning domains, signal sequences, and lipomodification signals), and the identification of several dysfunctional genes, including a remnant of a gene for anaerobic coproporphyrinogen III oxidase (*hemN*-like

gene), and a dysfunctional ABC-type transporter. The chapters II and III have been published, albeit in a somewhat different form (Noback *et al.*, 1996; Noback *et al.*, 1998).

In chapter four, the complete genome sequence of the *B. subtilis* genome is presented. This chapter has already been published (Kunst *et al*,. 1998)

Chapter five deals with the positional analysis of amino acid frequencies in fifteen known genomes. Comparison of average amino acid frequencies in the entire proteome with the average frequencies in the N- and C-terminal amino acids of these proteomes has revealed biases of many amino acids. Further investigation of the deduced (average) properties of the proteomes with respect to charge and hydrophobicity, showed that all proteomes display similar differences between the N- and C-termini with respect to these parameters. This could reflect that the N- and C-termini of proteins are usually located at the surface of proteins, and might be involved in the proper translocation of the nascent proteins through the ribosome.

In chapter six, the identification of the *gtaC* gene is presented. An ORF was found in the *prkA-addAB* region, *yhxB*, that displayed high similarity to phosphogluco- and phosphomanno-mutases. It was demonstrated that this ORF corresponds to the genetic marker *gtaC*, and is responsible for the glucosylation of cell wall teichoic acid and phage susceptibility in *B. subtilis*.

In chapter seven, functional analyses of the ubiquitous *hit* gene are presented. This ORF has highly conserved orthologs in (probably) all living organisms, including the organism with the smallest known genome, *M. genitalium*. Surprisingly though, the *hit* gene was not essential in *B. subtilis*, as shown by the viability of a strain with an insertional mutation of this gene. The insertional mutation of the *hit* gene induced a sensitive phenotype in heat-shock treatment. The Hit protein was found to have a twofold biochemical activity on ADP, the relative amounts of products of the Hit-mediated reaction being dependent on the pH. Hit hydrolyses ADP to AMP and Pi, and also acts as phosphotransferase in the reaction 2 ADP $\rightarrow$ ATP + AMP. The isolation of a transcriptional regulator of *hit*, the -also ubiquitous- *yabJ* gene, is also presented.

In chapter eight, the characterisation of a new forespore-specific gene of *B. subtilis*, *yhcN* is presented. Based on conserved amino acid sequence elements, specific for small acid soluble proteins (SASP; KLEVADE) and membrane anchored lipoproteins (LMTGC), this ORF was considered a likely candidate for involvement in spore formation. This assumption was shown to be correct; *yhcN* expression is dependent on the forespore-specific sigma factor σ^G, it is expressed at a very high level in the forespore and is located in the inner spore membrane. *YhcN* mutant spores show a phenotype of slower outgrowth than wild-type spores. This work has been published (Bagyan *et al.*, 1997).

Chapter nine is a typical example of paralog research. The *B. subtilis* genome encodes four closely related Type I signal peptidases, responsible for the removal of signal peptides from secretory precursor proteins, with overlapping substrate specificities and differing expression patterns. This work has been published (Tjalsma *et al.*, 1997)

References

Anagnostopoulos, C., Piggot, P. J., & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. *In Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, pp. 425-461. Edited by A. L. Sonenshein, J. A. Hoch, and R. Losick. Washington, DC: American Society for Microbiology.

Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., Näslund, A. K., Eriksson, A.-S., Winkler, H. H., & Kurland, C. G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396** , 133-140.

Bagyan, I., Noback, M., Bron, S., & Setlow, P. (1997). Characterization of *yhcN*, a new forespore-specific gene of *Bacillus subtilis*. *Gene* **212**, 179-188.

Bains, W. (1992). Sequence - so what? *Biotechnology* **10**, 751-761.

Bernhardt, J., Volker, U., Volker, A., Antelmann, H., Schmid, R., Mach, H., & Hecker, M. (1997). Specific and general stress proteins in *Bacillus subtilis* - A two dimensional protein electrophoretic study. *Microbiol.* **143**, 999-1017.

Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., & Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073.

Burland, V., Daniels, D. L., Plunkett, G., III, & Blattner, F. R. (1993). Genome sequencing on both strands: The Janus strategy. *Nucleic Acids Res.* **21**, 3385-3390.

Chee, M., Yang, D., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., & Fodor, S. P. A. (1996). Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-614.

Cheng, S., Chang, S. Y., Gravitt, P., & Respess, R. (1994). Long PCR. *Nature* **369**, 684-685.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., & Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., & Walters, L. (1998). New goals for the U.S. human genome project: 1998-2003. *Science* **282**, 682-689.

de Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W., & Mous, J. (1998). Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotech* **16**, 45-48.

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, J. M., Olsen, G. J., & Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353-358.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.

Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W. K., Koop, B., Hood, L., & Crkvenjakov, R. (1993). DNA sequence determination by hybridization: A strategy for efficient large-scale sequencing. *Science* **260**, 1649-1652.

Drmanac, S., Kita, D., Labat, I., Hauser, B., Schmidt, C., Burczak, J. D., & Drmanac, R. (1998). Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nature Biotech* **16**, 54-58.

Dujon, B. (1998). European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. *Electrophoresis* **19**, 617-624.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., Vugt, R. v., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., & Venter, J. C. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580-586.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A., & Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403.

Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren, E., Hardham, J. M., McLeod, M. P., Salzberg, S., Peterson, J., Khalak, H. G., Richardson, D., Howell, J. K., Chidambaran, M., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Cotton, M. D., Fujii, C., Garland, S., Hatch, B., Horst, K., Roberts, K., Sandusky, M., Weidman, J., Smith, H. O., & Venter, J. C. (1998). Complete genome sequence of *Treponema pallidum*, the Syphilis spirochete. *Science* **281**, 375-388.

Gas, S., Samson, F., Biaudet, V., Ehrlich, S. D., & Bessieres, P. (1998). Collecting functional analysis data in Micado. *In* The sixth European *Bacillus subtilis* gene function analysis meeting, pp. 7-7.

Gerstein, M. & Hegyi, H. (1998). Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiology Reviews* **22**, 277-304.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**, 546-567.

Griffin, H. G. & Griffin, A. M. (1993). DNA sequencing: Recent innovations and future trends. *Appl.Biochem.Biotechnol.* **38**, 147-159.

Guerdoux-Jamet, P., Hénaut, A., Nitschké, P., Risler, J.-L., & Danchin, A. (1997). Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? *DNA Res.* **4**, 257-265.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I., & Danchin, A. (1996). Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J.Mol.Biol.* **257**, 574-585.

Hieter, P. & Boguski, M. (1997). Functional genomics: it's all how you read it. *Science* **278**, 601-602.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420-4449.

Huynen, M. A. (1997). Differential genome display. *Trends Genet* **13** , 389-390.

Huynen, M. A. & Nimwegen, E. v. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* **15**, 583-589.

Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J.Mol.Biol.* **220**, 631-648.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., & Tabata, S. (1996). Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3** , 109-136.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Yoshizawa, T., Nakamura, Y., Robb, F. T., Horikoshi, K., Masuchi, Y., Shizuya, H., & Kikuchi, H. (1998). Complete sequence and gene organisation of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus hotikoshii* OT3. *DNA Res.* **5**, 55-76.

Klenk, H.-P., Clayton, R. A., Tomb, J.-F., White, O., Nelson, K., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, A. R., Graham, D. E., Kyrpides, N. C., Fleischmann, R. D., Quackenbush, J., Lee, N. H., Sutton, G. G., Gill, S., Kirkness, E. F., Dougherty, B. A., McKenney, K., Adams, M. D., Loftus, B., Peterson, S., Reich, C. I., McNeil, L. K., Badger, J. H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J. D., Weidman, J. F., McDonald, L., Utterback, T., Cotton, M. D., Spriggs, T., Artiach, P., Kaine, B. P., Sykes, S. M., Sadow, P. W., D'Andrea, K. P., Bowman, C., Fujii, C., Garland, S. A., Mason, T. M., Olsen, G. J., Fraser, C. M., Smith, H. O., Woese, C. R., & Venter, J. C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364-370.

Koonin, E. V. & Mushegian, A. R. (1996). Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr Opin Genet Dev* **6**, 757-762.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Düsterhöft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, H., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., & Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., & Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech* **14**, 1675-1680.

Marshall, A. & Hodgson, J. (1998). DNA chips: an array of possibilities. *Nature Biotech* **16**, 27-31.

Maxam, A. M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc.Natl.Acad.Sci.USA* **74**, 560-564.

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D., & Koornneef, M. (1998). *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282**, 662-682.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F., & Zollner, A. (1997). Overview of the yeast genome. *Nature* **387**, 7-9.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., & Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J.Mol.Biol.* **222**, 851-856.

Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete microbial genomes. *Proc.Natl.Acad.Sci.USA* **93**, 10268-10273.

Noback, M. A., Holsappel, S., Kiewiet, R., Terpstra, P., Wambutt, R., Wedler, H., Venema, G., & Bron, S. (1998). The 172 kb *prkA-addAB* region from 83° to 97° of the *Bacillus subtilis* chromosome contains several dysfunctional genes, the *glyB* marker, many genes encoding transporter proteins, and the ubiquitous *hit* gene. *Microbiol.* **144**, 859-875.

Noback, M. A., Terpstra, P., Holsappel, S., Venema, G., & Bron, S. (1996). A 22 kb DNA sequence in the *cspB-glpPFKD* region at 75° on the *Bacillus subtilis* chromosome. *Microbiol.* **142**,

Phizicky, E. M. & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol.Rev.* **59**, 94-123.

Ramsay, G. (1998). DNA chips: state-of-the art. *Nature Biotech* **16**, 40-44.

Rocha, E. P. C., Viari, A., & Danchin, A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res* **26**, 2971-2980.

Rosa, R. d. & Labedan, B. (1998). The evolutionary relationships between the two bacteria *Eschrichia coli* and *Haemophilus influenzae* and their putative last common ancestor. *Mol Biol Evol* **15**, 17-27.

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotech* **16**, 939-945.

Sanger, F., Nicklen, S., & Coulsen, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.USA* **74**, 5467

Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E., & Davis, R. W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *TIBTech.* **16**, 301-306.

Schmid, R., Bernhardt, J., Antelmann, H., Volker, A., Mach, H., Volker, U., & Hecker, M. (1997). Identification of vegetative proteins for a two-dimensional protein index of *Bacillus subtilis*. *Microbiol.* **143**, 991-998.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G. M., Daniels, C. J., Mao, J., Rice, P., Nölling, J., & Reeve, J. N. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. *J.Bacteriol.* **179**, 7135-7155.

Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., Koonin, E. V., & Davis, R. W. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754-759.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.

The C.elegans sequencing consortium (see genome.wustl.edu/gsc/C_elegans) (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012-2018.

Tjalsma, H., Noback, M. A., Bron, S., Venema, G., Yamane, K., & Van Dijl, J. M. (1997). *Bacillus subtilis* contains four closely related type I signal peptidases with overlapping substrate specificities. Constitutive and temporally controlled expression of different *sip* genes. *J.Biol.Chem.* **272**, 25983-25992.

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H.-P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., FitzGerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547.

Vagner, V., Dervyn, E., & Ehrlich, S. D. (1998). A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiol.* **144**, 3097-3104.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270**, 484-487.

Wallin, E. & Von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 1029-1038.

Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H., & Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotech* **15**, 1359-1367.

CHAPTER II

Sequencing and annotation of the 172 kb DNA region from 83° to 97° of the *Bacillus subtilis* chromosome

II.1. Summary

A 171,812 bp nucleotide sequence between *prkA* and *addAB* (83° to 97º) on the genetic map of the *Bacillus subtilis* 168 chromosome was determined and analyzed. An accurate physical/genetic map of this previously poorly described chromosomal region was constructed. Hundred-seventy open reading frames (ORFs) were identified on this DNA fragment. These include the previously described genes *cspB, glpPFKD, spoVR, phoAIV, papQ, citRA, sspB, prsA, hpr, pbpF, hemEHY, aprE, comK,* and *addAB*. ORF *yhaF* in this region corresponds to the *glyB* marker. Among the striking features of this region are: an abundance of genes encoding (putative) transporter proteins, several dysfunctional genes, the ubiquitous *hit* gene, and five multidrug-resistance-like genes.

Accession numbers: EMBL accession numbers for the sequences reported in this paper are X96983, Y14077, Y14078, Y14079, Y14080, Y14081, Y14082, Y14083, & Y14084

II.2. Introduction

Since 1995, when the Gram-negative bacterium *Haemophilus influenzae* was the first free-living organism to be entirely determined at the DNA level (Fleischmann *et al.*, 1995), the sequences of several other genomes were elucidated. Among these are the smallest known genome of the bacterium *Mycoplasma genitalium* (Fraser *et al.*, 1995), the genome of the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996), the bacterium *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), the bacterium *Escherichia coli* (O'Brien, 1997), the cyanobacterium *Synechocystis* PCC6803 (Kaneko *et al.*, 1996), and the eukaryote *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996; Mewes *et al.*, 1997). In the framework of the combined European/Japanese *Bacillus subtilis* genome sequencing project that was recently completed (Kunst *et al.*, 1997), a 171,812 bp DNA sequence, representing 4.1 % of the genome, was determined and analysed in our group. The sequence spans the region between 83° (*prkA*) and 97° (*addAB*) on the genetic map of the *B. subtilis* chromosome (Anagnostopoulos *et al.*, 1993; Biaudet *et al.*, 1996). The present paper deals with the cloning,

sequencing and *in silico* analysis of putative genes in this region. A correction of the existing genetic (Anagnostopoulos *et al.*, 1993; Biaudet *et al.*, 1996) and physical maps (Itaya & Tanaka, 1991) is presented.

II.3. Methods

Bacterial strains and DNA handling procedures

B. subtilis 168 (*trpC2*) was used as the standard strain for sequence determinations. DNA fragments for sequencing were obtained mainly by Long-Range PCR (LR PCR; Cheng *et al.*, 1994; Barnes, 1994), or inverse Long-Range PCR (i-LR PCR) techniques, using the Gene Amp XL-PCR kit with rTth polymerase of Perkin Elmer (Norwalk, CT, USA). All amplification reactions were performed according to the protocols supplied by the manufacturer. I-LR PCR was performed by digestion of *B. subtilis* chromosomal DNA with appropriate restriction enzymes, followed by purification of the digested DNA, and subsequent self-ligation at low concentrations of DNA (<5 μg ml^{-1}). PCR primers used are listed in Table II.1. An overview of the amplified fragments is presented in Fig. II.1.

Some fragments were cloned as phage lambda DNA inserts. The *B. subtilis* lambda EMBL12 library, constructed from a sized partial *Sau*3A digest of the chromosome (kindly provided by Dr. C. Harwood and A. Wipat, Newcastle upon Tyne, UK), was screened by the plaque hybridisation method (Sambrook *et al.*, 1989) for the presence of desired sequences. For sequence determinations, phage lambda DNA inserts were amplified by LR PCR and subsequently processed in the same way as other LR PCR fragments (see below).

PCR fragments used for sequencing were treated in one of the following ways:

i) Shotgun cloning in M13mp18 phage by *nebulization*. Amplified DNA fragments were sheared by nebulization under nitrogen-gas pressure using a DNA Nebulizer (obtained from GATC GmbH, Konstanz, Germany), according to the instructions of the supplier. The sheared DNA was treated with Klenow enzyme (Boehringer, GmbH, Mannheim, Germany), in the presence of a mixture of the four deoxyribonucleotides (dNTPs). The DNA mixture was fractionated according to size by agarose gel electrophoresis, and segments in the 500-1000 bp range were extracted from the agarose using the JETsorb DNA extraction kit (GENOMED GmbH, Oeynhausen, Germany). The DNA fraction obtained was treated with T4 DNA polymerase and dNTPs (Boehringer) to obtain blunt-ended fragments. This DNA mixture was ligated into the M13mp18 phage vector, which had been digested with *SmaI* and treated with alkaline phosphatase (Boehringer), and the ligation mixture was used to transform the *Escherichia coli* strain XL1-Blue (*supE*$^{+}$ *lac*$^{-}$ *hsdR17 recA1 [F' proAB*$^{+}$ *lacI*^{q} *lacZ**D***M15]).

ii) Shotgun cloning in pUC18 after *limited DNaseI digestion* in buffer consisting of 500 mM Tris-HCl, pH7.6; 100 mM $MnCl_2$; 1 mg ml^{-1} BSA. Subsequently, the DNA fragments were treated with T4 DNA polymerase and Klenow enzyme (in 10 mM Tris-HCl, pH 8.5; 0.25 mM dNTPs; 5 mM $MgCl_2$) and fractionated by agarose gel electrophoresis. Fragments ranging from 500 to1500 bp were extracted and ligated into pUC18 which had been digested with *Sma*I and treated with alkaline phosphatase. The ligation mixtures were used to transform the *E. coli* strain XL1-Blue (*supE*$^{+}$ *lac*$^{-}$ *hsdR17 recA1* [F' *proAB*$^{+}$ *lacI*^{q} *lacZ**D***M15])

(Stratagene, La Jolla, CA, USA). DNA inserts were sequenced by the method described below.

iii) Sequencing directly on PCR-generated DNA. To prevent sequencing mistakes that were generated during the PCR reaction, eight separate amplification reactions were performed and the products were pooled.

Table II.1. Primer sequences, their position in the sequenced region, type of amplification, and the second primer that was used in the amplification procedure.

PRIMER	SEQUENCE (5'->3')	POSITION	AMPLIFICATION
SH25	CGG TAT ATA TCT GGC GGA GCT GCA T	29268 C	+ xlp02 LR PCR
xlp01b	TGT AAC GGT TGT CAA AGA ACA GGA AC	35832	+ xlp21 i-LR PCR *SspI*
xlp02	CTA GTG ATC GCA GGC TAT GGA GGC T	23377	+ SH25 LR PCR
xlp03	GCA GGT CGT CAG AAT CAG CTC TTC C	23868 C	+ xlp10 LR PCR
xlp04	GTA TAC CGA ACA GCG TGG CTC AGA A	145844	+ xlp08 LR PCR
xlp05	CCT GTT CGG TCA GCT CCT TCC TAT T	146021	+ xlp07 LR PCR
xlp06	CGG CTC TTC ACT CTC AAG GCT ACA C	133516 C	+ xlp36 LR PCR
xlp07	CTG TAG AAC CAG TAG GTC CGC CAA G	133157	+ xlp05 LR PCR
xlp08	GCT GAT TAT CTC CGC ACA TCT CTC C	164524 C	+ xlp04 LR PCR
xlp09	GTC ATA TTC GGC TCT AGC TTC CTG C	18726 C	+ xlp11 i-LR PCR *SalI*
xlp10	CTG ATC GAG ACT GGC AGG AAG C	18689	+ xlp03 LR PCR
xlp11	CTG TTC CAT ATC CTG CGC ATC AAG	19030	+ xlp09 i-LR PCR *SalI*
xlp12	GAA GCC TTC GCC TTG AAT AGC AGA G	12695	+ xlp13 i-LR PCR *AsuII*
xlp13	TGC CAT CCA CAT ACT GAG TCA AGT C	12397 C	+ xlp12 i-LR PCR *AsuII*
xlp17	GGT GAC AGC CTC AAT CGT ATC CAT C	90063 C	+ xlp18 i-LR PCR *PstI*
xlp18	GAA GGA CCA AGG ATC ACC AAG AAG G	90500	+ xlp17 i-LR PCR *PstI*
xlp20	GGA TCG ACA GAC TTG GCT ACT TGT G	7947	+ xlp28 i-LR PCR *EcoRI*
xlp21	GCT TCC TCA CCT TGC TTC GAG ATG T	35360 C	+ xlp1b i-LR PCR *SspI*
xlp28	GAC ATT GGA ATC GAG TGA TGC GTG	7557 C	+ xlp20 i-LR PCR *EcoRI*
xlp35	GAT GAT CCC GCT GAA AGA GTT GAG G	79421 C	+ LT7 LR-PCR on λ
xlp36	AGA ATA GTT CCG AGC GGC TCA GTT G	109109	+ xlp06 LR PCR
xlp38	GCA CAT GTT TTA AGC CGC AAA CCG	41808	+ LT7 LR PCR on λ
xlp401	GAC GAT GAA TTG TTT ACT CCG ACC	50328	+ xlp402 LR PCR
xlp402	GCG CAC TTG GTG TTC CAG TCA TAG	71296 C	+ xlp401 LR PCR
LT7	GCC TAA TAC GAC TCA CTA TAG GGA G		λGEM-11 Left arm
LSP6	GGC CAT TTA GGT GAC ACT ATA GAA G		λGEM-11 Right arm

The column 'Position' represents the position of the primer in the *prkA-addAB* region with respect to the first base, in gene *yzdA*. A capital C means that the primer is on the complementary strand. In the column 'Amplification', the second primer used for the amplification is indicated. In the case of i-LR PCR, the restriction enzyme that was used for digestion of the chromosome is also specified. The addition of 'λPCR' means that the insert of a recombinant lambda phage was amplified.

Sequence determination

DNA was isolated on the Vistra DNA Labstation 625 (Amersham, Rainham, UK) using either the "automated M13 template preparation kit" or the "automated plasmid preparation kit". DNA inserts were sequenced by the dideoxy chain termination method (Sanger *et al.*, 1977) using the Amersham "automated Delta taq cycle sequencing kit" and the Amersham Vistra automated DNA sequencer 725. The universal forward sequencing primer was used (5'GTAAAACGACGGCCAGT3'). Remaining gaps between the contiguous sequences obtained through shotgun cloning were determined by primer walking on PCR material using the Amersham "sequenase PCR product sequencing kit" and [^{35}S]-dATP.

Data handling and computer analysis

DNA sequences were assembled using the Staden package (Dear & Staden, 1991); obtained from MRC, Cambridge, UK). A redundancy of at least four readings per base, with a minimum of one reading for each strand, was taken as a standard for a reliable sequence. The compiled sequence was analyzed for the presence of ORFs consisting of more than 50 codons using the Staden package. The amino acid (a.a.) sequences of the putative protein products encoded by the ORFs were analyzed for similarities to known sequences in databanks using the FASTA program (Pearson & Lipman, 1988), and the BLAST E-mail server at the NCBI (retrieve@ncbi.nlm.nih.gov).

Transformation and competence

B. subtilis cells were made competent essentially as described by Bron and Venema (1972). *E. coli* cells were made competent and transformed by the method of Mandel and Higa (1970).

Isolation of DNA

B. subtilis chromosomal DNA was purified as described by Bron (1990). Plasmid DNA was isolated by the alkaline-lysis method of Ish-Horowicz and Burke (1981).

II.4. Results and discussion

Cloning of the *prkA-addAB* region

For the cloning of the *prkA-addAB* region we started from two marker regions on the genetic map: the *glpPFKD* operon, which was already cloned and sequenced (Beijer *et al.*, 1993; Holmberg *et al.*, 1990) and the *glyB* marker, which was only genetically mapped (Harford *et al.*, 1976). The cloning and analysis of the *yhcA* to *glpP* region (22 kb), which is part of the *prkA-addAB* region, has been reported in a previous paper (Noback *et al.*, 1996).

An overview of cloned fragments from this region, and the method by which they were obtained, is shown in Fig. II.1. Fragments indicated in this figure as ‘formerly known’ were partially (at least 10 % of their length) resequenced. Other previously known sequences (*cspB, sspB, prsA, hpr, hemEHY, aprE* & *comK*) were resequenced in their entirety. In a total of about 15 kb of resequenced DNA, less than ten discrepancies were found, and these were all present in non-coding areas.

By i-LR PCR, using *Eco*RI from *yhcA* outward in the direction of *prkA* (Fischer *et al.*, 1996), a 5 kb fragment was amplified which spans the region from *yzdC* to *yhcA*. In the other direction, from *glpD* outward in the direction of *addAB*, an i-LR PCR fragment of 7 kb was obtained using *Ssp*I and primers XLP21 and XLP1B. Using a terminal part of this fragment as probe, a lambda DNA clone was isolated containing an additional 3 kb. This fragment unexpectedly proved to contain part of the *spoVR-citA* contig (Beall & Moran Jr, 1994; Hulett *et al.*, 1991; Jin & Sonenshein, 1994), which was already present in SubtiList (the central database for *B. subtilis* sequences (Moszer *et al.*, 1995), and had previously been mapped outside our region.

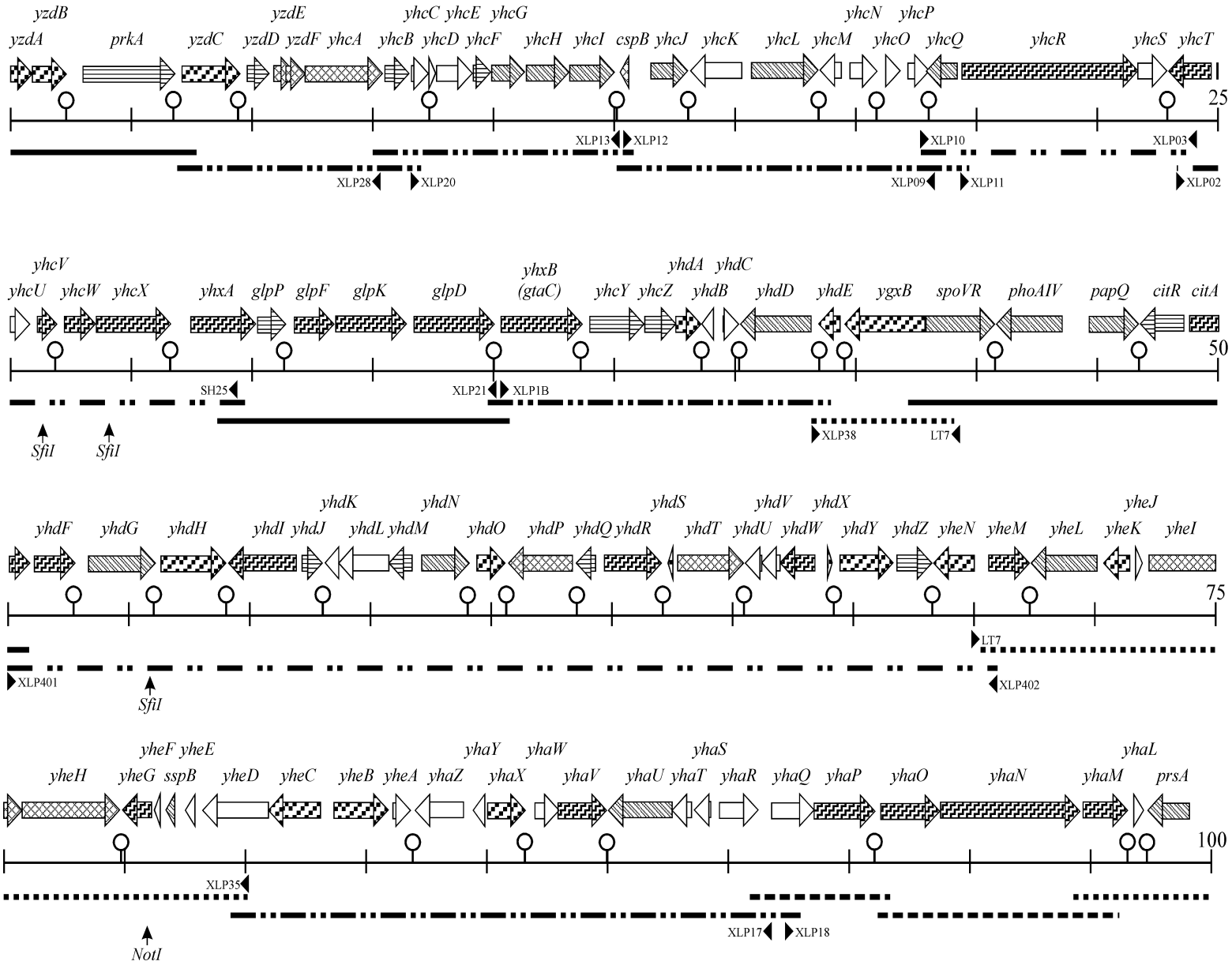
A 13.5 kb clone was isolated by screening a lambda-GEM11 genome bank with a 4.5 kb *glyB*$^+$ *Sac*I chromosomal fragment (kindly provided by Dr. M. Sarvas, Helsinki, Finland). Southern analysis revealed that this clone also contained the *hpr* (Perego & Hoch, 1988) and *prsA* (Kontinen *et al.*, 1991) genes. By plasmid rescue, 'walking' in the direction of *prkA*, two *E. coli* plasmid clones were isolated containing *yhaO-yhaM* (5 kb) and *yhaR-yhaP* (4 kb), respectively. Using the divergent primers XLP17 and XLP18, and *PstI*-digested chromosomal DNA, a 12 kb DNA fragment was amplified by i-LR PCR (*yhaR* to *yheD*). Using the *yheD* end of this fragment as a probe, a clone was isolated from a lambda-GEM11 genomic bank that contained the *yheD-yheM* region (9 kb). Using a primer from the end of this clone, XLP402, we were able to amplify the region between *yheM* and *citA* (primer XLP401) by LR PCR, yielding a fragment of 21 kb.

Finally, three LR PCR fragments were obtained which together span the region between *glyB* and *addAB*. First, a 26 kb fragment between *yhaA* and *aprE* (Stahl & Ferrari, 1984) was amplified using primers XLP36 & XLP06. Unexpectedly, this fragment contained the *hemEHY* gene cluster (Hansson & Hederstedt, 1992) that was formerly mapped at a different position (94 degrees). Second, a 12.5 kb fragment was generated between *aprE* and *comK* (primers XLP07 & XLP05). Finally, a PCR fragment was obtained between *comK* and *addB* (primers XLP04 & XLP08), yielding a fragment of 18 kb.

Assignment of ORFs

ORFs were searched in all six possible reading frames and selected according to the following criteria. A putative ORF should have an ATG, TTG or GTG start codon preceded within 5-15 bp by a Shine-Dalgarno (SD) sequence that is (partly) complementary to the 3' end of the *B. subtilis* 16S rRNA (3' UCUUUCCUCCACUAG 5'). We also selected ORFs on the basis of codon usage statistics, using the Bsu.cod table on the EMBL CD-ROM. In total, hundred seventy open reading frames were identified, and these are indicated in Fig. II.1. The protein coding density of this region is 90 %. Fifty-eight percent of the putative ORFs is transcribed in the direction of replication fork movement (clockwise); fourty-two percent is transcribed in the counterclockwise direction. Seventy-three percent of the ORFs have an ATG start codon; 15% TTG, and 9 percent GTG. One ORF, *yhdQ*, putatively has the rare ATA start codon. The average size of the ORFs from this region is 302 a.a. The classification of these ORFs according to their putative function (also indicated in Fig. II.1.) is described in the following paragraph.

In Table II.2, the coordinates of the ORFs relative to the first base in this region, their size in a.a., the calculated molecular masses (kDa) and isoelectric points (pIs), and the putative Shine-Dalgarno sequences are listed. The nomenclature of the ORFs is according to agreements made among the participants in the European/Japanese *B. subtilis* genome sequencing project.


yzdB
yzdA
prkA
yzdC
yzdE
yzdD yzdF yhcA
yhcC yhcE yhcG
yhcB yhcD yhcF
yhcH
yhcI
cspB yhcJ
yhcK
yhcL
yhcN yhcP
yhcM
yhcO
yhcQ
yhcR
yhcS yhcT
25
XLP13
XLP12
XLP10
XLP03
XLP28
XLP20
XLP09
XLP11
XLP02
yhcV
yhcU yhcW yhcX
yhxA
glpP glpF glpK
glpD
yhxB
(gtaC)
yhcY yhcZ
yhdA yhdC
yhdB
yhdD
yhdE
ygxB
spoVR
phoAIV
papQ
citR citA
50
SH25
XLP21
XLP1B
SfiI
SfiI
XLP38
LT7
yhdK
yhdN
yhdS
yhdV yhdX
yheJ
yhdF
yhdG
yhdH
yhdI yhdJ
yhdL yhdM
yhdO
yhdP
yhdQ yhdR
yhdT
yhdU yhdW
yhdY
yhdZ
yheN
yheM
yheL
yheK
yheI
75
LT7
XLP401
XLP402
SfiI
yheF yheE
yhaY
yhaW
yhaS
yhaL
yheH
yheG sspB
yheD
yheC
yheB yheA yhaZ
yhaX
yhaV
yhaU yhaT
yhaR
yhaQ
yhaP
yhaO
yhaN
yhaM
prsA
100
XLP35
XLP17
XLP18
NotI

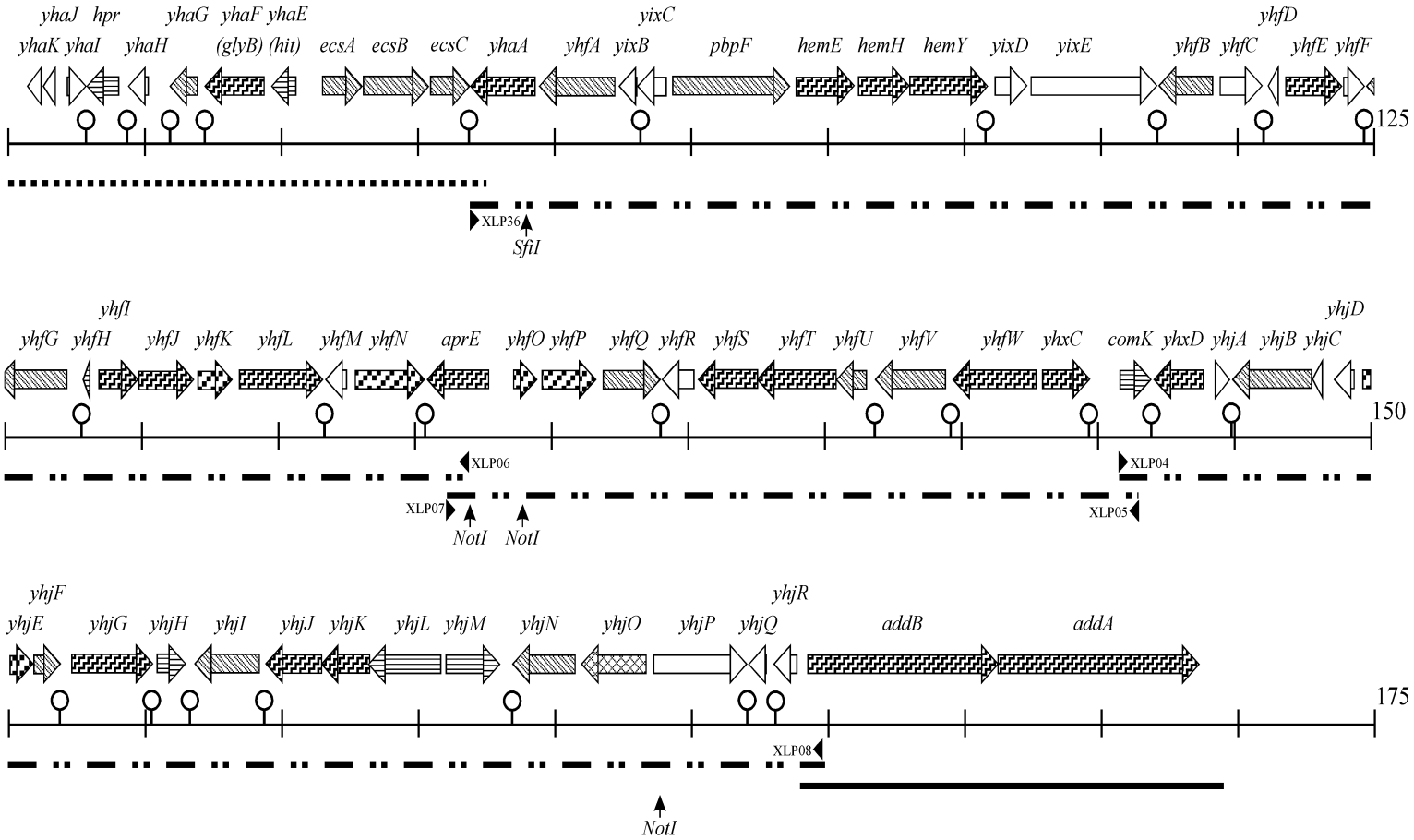


Fig. II.1. Overview of the *prkA-addAB* region. Below the line representing the map, which is divided into 25 kb fragments, *Not*I and *Sfi*I restriction sites and the method used to obtain the clones are indicated: ——— formerly known sequences spanning more than 5 kb; - - - - plasmid rescue clones; ····· fragments cloned in lambda phage; — ·· — fragments cloned by linear LR PCR; —·—·— fragments cloned by i-LR PCR. Above the line, the ORFs, their classification, and the presence of terminator-like sequences are indicated. The ORFs were classified as follows: ▭ genes of unknown function without homologues in public databases; ▨ genes of unknown function with unknown homologues in public databases; ▤ genes involved in information pathways; ▩ genes involved in intermediary metabolism; ▧ genes for cell envelope and cellular processes; ▨ other. ⚲ Terminator-like sequence. See text for further details.

Table II.2. Co-ordinates of ORFs within the *prkA-addAB* region, their size in amino acids and mass (kD), calculated pI, putative S-D sequence, and initiation codon

ORF	Endpoints (nucleotides)	Size of deduced product		Calcu-lated	SD consensus sequence (uppercase) and initiation codon (bold)
		# a.a.	mass (kD)	pI	
yzdA	1->431	141	15.2	5.09	N.P.
yzdB	443->1150	234	25.1	6.90	GtAAGGAGGatcgta**ATG**
prkA	1500->3395	630	72.9	6.36	AtAgAGGAGGTccTt**ATG**
yzdC	3575->4753	391	45.3	5.63	AAGGAGGgGAattc**ATG**
yzdD	4913->5377	154	17.6	8.56	AGgAAtGAGGTGAaaaggag**TTG**
yzdE	5446->5850	134	14.9	9.64	GAAAGGAGaaaAcaaa**ATG**
yzdF	5697->6077	126	13.7	4.49	N.P.
yhcA	6118->7716	532	58.3	9.30	GAAAGGAGGTtgTCttag**ATG**
yhcB	7739->8269	176	19.0	4.74	AgGGGGTttcCtga**ATG**
yhcC	8282->8656	124	14.0	5.43	aAggaGAGGTtgaa**ATG**
yhcD	8656->8811	51	6.0	9.75	AGAAAaagta**ATG**
yhcE	8816->9577	253	29.5	9.47	GGAGGTaAagac**ATG**
yhcF	9580->9945	121	14.0	5.59	aGAGGTGtaaat**ATG**
yhcG	9947->10645	232	26.5	5.52	agAgGGAGGctAaa**ATG**
yhcH	10662->11579	305	34.5	6.63	AaAgAGGAGGaatatg**ATG**
yhcI	11572->12513	313	34.9	6.61	AaAgAGGAGGTtcagc**ATG**
cspB	12605<-12808	67	7.4	4.47	AGGAGGaaATttc**ATG**
yhcJ	13244->14035	263	29.2	5.21	AGGAGtatggtcaca**ATG**
yhcK	14076<-15155	359	40.7	8.56	aagGGTGATaatat**TTG**
yhcL	15328->16716	463	49.0	9.14	GAAgGGAGagtttacctgct**TTG**
yhcM	16759<-17214	151	17.0	9.55	AAAGGAGGgatc**ATG**
yhcN	17364->17933	189	21.0	5.44	AAAGGAGGaatTCac**ATG**
yhcO	18113->18412	99	11.4	9.56	GGAGtccttgtg**ATG**
yhcP	18403->19020	205	24.1	4.94	GGAGGcttaCtccggttta**TTG**
yhcQ	18952<-19605	217	24.8	6.00	AAAGGAGGaatTCggt**TTG**
yhcR	19688->23341	1217	132.7	4.79	GAAAGGAatTat**ATG**
yhcS	23338->23934	198	22.9	7.31	AAAGGAGcgccTCcagaac**GTG**
yhcT	23964<-24872	302	33.7	9.28	AAAGGAGccatTtaac**ATG**
yhcU	24983->25375	131	15.3	8.99	AGGAtaTtcg**ATG**
yhcV	25515->25937	140	14.9	5.13	GAAAGGgGtgctgaca**ATG**
yhcW	26064->26726	220	24.6	4.74	AAAGGAGtTGtaCcca**GTG**
yhcX	26742->28283	513	60.2	5.51	AGAAAGGAGcgagTagg**TTG**
yhxA	28703->30055	450	49.9	5.87	AGGgaacGcTaatgaa**ATG**
glpP	30083->30661	192	21.6	8.08	AAAGGAGcac**ATG**
glpF	30840->31664	274	28.7	9.30	AGGAGGaatgtgct**ATG**
glpK	31683->33173	496	55.1	5.10	AAaGgGGaGAcatctt**ATG**
glpD	33314->34981	555	62.5	7.96	AacAAGGAGGaaAcgta**ATG**
yhxB	35113->36810	565	62.9	5.03	AcAtAGGAGGacgaat**ATG**
yhcY	36959->38098	379	42.0	6.90	GGAGtgagaaac**GTG**
yhcZ	38095->38739	214	24.0	6.16	AAAGGAGGgGcggt**ATG**
yhdA	38736->39260	174	18.9	6.77	gAAtGgaGaGATCtcaaa**ATG**
yhdB	39275<-39517	80	9.8	4.68	AGAAAGGAGaaGcgattc**ATG**
yhdC	39718->40041	107	12.3	6.59	AcAGGAGactgaaaaa**ATG**
yhdD	40082<-41548	488	51.4	10.03	AaAAAGGAGaactaag**ATG**
yhdE	41701<-42142	146	16.6	7.83	GAGGTctTatt**ATG**
ygxB	42244<-43902	552	60.0	9.85	GGActTatctata**ATG**
spoVR	43933->45339	468	55.6	5.62	AgtaGgGGgGATtcgg**TTG**
phoAIV	45369<-46754	461	50.3	9.52	AAAGGAGGc**ATG**aaaaaa**ATG**
papQ	47286->48317	343	37.3	10.31	GGAGGaaAat**ATG**
citR	48336<-49262	308	35.6	8.64	AgGGAGaatAgaa**ATG**
citA	49371->50471	366	40.9	6.03	GAtAGGaGGaataCaa**ATG**
yhdF	50545->51414	289	31.5	5.31	AGGAGtgatgaat**GTG**
yhdG	51664->53061	465	49.7	9.46	GGAGtTGAaggggga**ATG**

yhdH	53179->54534	451	48.9	9.48	GAAAGGAaGTGAcgttta**TTG**
yhdI	54569<-55978	469	52.8	7.07	AcAAAGGAGacatgag**ATG**
yhdJ	56088->56516	142	16.4	8.26	GAAAGGgGaTtgagaag**ATG**
yhdK	56547<-56837	96	10.6	8.20	GcgAGGtGGaatT**ATG**
yhdL	56825<-57901	358	40.6	6.42	AtAAtaGAGGTGtTa**ATG**
yhdM	57891<-58382	163	19.4	7.04	AgaGgGGaGAaaaggca**GTG**
yhdN	58579->59574	331	37.3	4.87	AAGGAGtgGcaCa**ATG**
yhdO	59709->60308	199	21.9	9.58	AcAAAGGAaGTGcgat**ATG**
yhdP	60377<-61711	444	49.8	4.50	AGAgtGaAGGTtcTaa**TTG**
yhdQ	61772<-62188	138	15.8	7.95	GttgGGAGGgat**ATA**
yhdR	62360->63541	393	43.9	5.13	GGAaGgGAcag**ATG**
yhdS	63681<-63791	35	4.1	8.21	AacAttGAGGTacgCg**GTG**
yhdT	63868->65253	461	51.5	4.86	GttgaGAGGatAgggttaa**ATG**
yhdU	65267<-65623	118	12.4	9.52	AGAAAGGgGcTGcaggaaaa**ATG**
yhdV	65620<-66015	131	13.9	10.04	AtAAAGGAtGgcAaac**ATG**
yhdW	66002<-66733	243	27.5	9.24	AGGAGcTGActtagc**TTG**
yhdX	66967->67074	35	4.0	9.70	AaAAAGGAGGcGAgatc**ATG**
yhdY	67223->68338	371	42.5	5.98	AgaGgGGaGAcagtc**ATG**
yhdZ	68408->69151	247	27.4	5.28	AaAAAGGcGGTGtTgag**TTG**
yheN	69175<-70023	282	31.7	8.46	AAtGGAGagatTgtt**ATG**
yheM	70308->71156	282	31.2	4.99	AaAAgGGAGGgctTtt**ATG**
yheL	71199<-72560	453	48.0	7.98	AaAtAtGgGGTGtTatt**TTG**
yheK	72687<-73241	184	20.1	4.92	AtAGGAaaaGgTtaa**TTG**
yheJ	73351->73512	53	6.3	10.64	AAGGAatTtgc**GTG**
yheI	73632->75389	585	65.1	6.52	AGGAGaTGgggtag**ATG**
yheH	75386->77407	673	76.3	7.31	AagAAGGgGGaGcaggggc**ATG**
yheG	77456<-78076	206	22.8	6.04	GcgAGGAGGTttTtta**ATG**
yheF	78115<-78240	41	5.0	8.22	AaAAgGGAGGgaATCggg**GTG**
sspB	78345<-78548	67	7.0	4.87	AaAAAGGAGaTttTacac**ATG**
yheE	78757<-78975	72	8.5	6.17	GtAAGGAGc**GTG**
yheD	79125<-80486	453	51.4	8.61	AGAAAGGAGtTctTCcgc**GTG**
yheC	80476<-81567	363	41.9	9.03	AAAGGgaGaGtctcacc**ATG**
yheB	81834->82967	377	42.9	8.96	AaGGAGGaagatgaataggaa**ATG**
yheA	83060->83413	117	13.6	4.53	GAAAGGAGcTatTtaca**ATG**
yhaZ	83457<-84530	357	41.8	8.86	AAAaGcGGTGtTtat**ATG**
yhaY	84723<-84974	83	9.6	9.64	GtgtatAGGat**ATG**
yhaX	85017->85814	265	29.2	7.13	AaggAGGgGGacATCtct**GTG**
yhaW	85994->86494	166	19.0	6.34	AtAAAttAGGTGATgaag**TTG**
yhaV	86458->87498	346	39.7	6.07	N.P.
yhaU	87516<-88742	408	43.9	8.97	GGAGagGgcgt**GTG**
yhaT	88739<-89236	165	18.7	5.00	GGAGGgatTttca**TTG**
yhaS	89300<-89638	112	12.8	7.15	AaAAAGaAGGgatatcttg**ATG**
yhaR	89803->90630	275	29.5	6.13	AtGGAGGTGcTttta**ATG**
yhaQ	90901->91797	298	33.8	6.37	GcAAGcAGGaGATtca**GTG**
yhaP	91790->93049	419	45.4	5.42	AAAGGtGGgGgcCgtct**ATG**
yhaO	93156->94382	408	46.8	5.40	GAAAGGAGcaGA**atg**TTG
yhaN	94396->97278	963	111.1	6.03	AtAcAtGAGGcG**gTg**acagct**TTG**
yhaM	97352->98296	314	35.7	6.12	AcGGAGGgagctttaatagaa**ATG**
yhaL	98421->98633	70	8.4	5.08	AagggGGAGGaGccC**GTG**
prsA	98674<-99552	292	32.5	8.77	AGGAGtgttTgaaaaca**ATG**
yhaK	100352<-100612	86	9.7	8.93	AGAAAaaAaGTttTacata**TTG**
yhaJ	100630<-100869	79	8.9	8.66	AAGGAtGactTtg**ATG**
yhaI	101077->101418	113	13.3	4.36	AGAAAGaAGtgGtgtgg**ATG**
hpr	101415<-102026	203	23.7	5.34	AAGcAGGTGAcgta**ATG**
yhaH	102204<-102560	118	13.1	8.33	AaAAgacgGGTGATtgta**ATG**
yhaG	102953<-103471	172	18.3	10.70	AgAGGAGagcATagtt**ATG**
yhaF	103596<-104675	359	40.1	5.72	AaAcAGGgaGaGATCata**ATG**
yhaE	104825<-105259	145	16.3	6.41	AAGGAGGaaccCtc**ATG**
ecsA	105747->106490	247	27.7	5.86	AcAtAaGgGGaGAaact**ATG**

ecsB	106483->107709	408	47.3	9.95	AcAAAGGAaGacgctggcc**ATG**
ecsC	107729->108439	236	26.7	8.77	GAAAaGAGGTaATCaa**ATG**
yhaA	108457<-109647	396	43.3	6.14	AAAGGgGGaagcggct**TTG**
yhfA	109720<-111111	463	48.8	7.47	AaGgaGTGATCg**ATG**
yixB	111177<-111491	104	12.0	9.57	GGAGGaaAgCaaa**ATG**
yixC	111536<-112036	166	18.8	5.38	AagcAGGAGGTGgcTgat**ATG**
pbpF	112158->114302	714	79.3	7.29	AaAggcGAGGTGAgttc**ATG**
hemE	114424->115485	353	39.7	5.40	GAAAGGtGGaaATCag**ATG**
hemH	115557->116489	310	35.3	4.72	AAAGagGGTGtaaaca**GTG**
hemY	116504->117916	470	51.2	8.08	AAAGaAGGcGATgaac**ATG**
yixD	118062->118637	191	21.8	7.38	AGtttcGAGGTGAataca**ATG**
yixE	118708->121035	775	84.1	5.08	AGAAtGGAGGcatcagg**ATG**
yhfB	121077<-122054	325	35.4	5.90	AAGGAGtgatTCat**ATG**
yhfC	122180->122956	258	28.7	8.73	AaAAAGGAGGctgaaaa**ATG**
yhfD	123047<-123250	67	8.5	8.11	AgAGGAGGgatTtct**ATG**
yhfE	123369->124409	346	38.7	6.16	AAAGGAGGaatgCcct**ATG**
yhfF	124422->124829	135	15.3	4.52	AAGGgGGaGgaCca**ATG**
yhfG	124866<-126155	429	45.9	8.82	GGAGGTaATCt**ATG**
yhfH	126426<-126560	43	5.1	6.92	GGgAAaGaGGgA**Ttg**gtt**ATG**
yhfI	126718->127452	244	26.5	5.86	AGAtAGGAGGacATt**ATG**
yhfJ	127465->128460	331	38.0	6.09	AtAAAGGAGGaGcaCc**ATG**
yhfK	128525->129169	214	22.8	5.30	AGgcAGGAGGgatTCac**ATG**
yhfL	129286->130827	513	56.6	5.46	ActtAaGgGGTGggaga**ATG**
yhfM	130866<-131261	131	15.0	8.25	GtttGGAGtgatgCaa**ATG**
yhfN	131410->132690	426	48.9	6.35	GtgAGGAGtgaggCgtt**ATG**
aprE	132729<-133874	381	39.5	9.08	AAAGGAGagGgTaaaga**GTG**
yhfO	134309->134758	149	16.7	7.99	AGggAGGAaGaaATaag**ATG**
yhfP	134830->135822	330	34.8	4.80	AAAGGAGtgtgcCga**ATG**
yhfQ	135964->137010	348	38.6	8.96	AaAtAattGGTGATa**ATG**
yhfR	137042<-137623	193	22.0	5.31	AgGaAGGgGATttt**ATG**
yhfS	137694<-138788	364	38.4	5.25	GgAAGagaGTGtaCagtataa**ATG**
yhfT	138785<-140224	479	52.9	6.20	AAAGGAGGatgaCaatac**ATG**
yhfU	140231<-140791	186	20.0	10.32	GGAGGattTCac**ATG**
yhfV	140926<-142224	432	48.8	5.62	AAGGgGGatcattgta**TTG**
yhfW	142363<-143892	509	57.1	5.90	GAttGGAGGTataacggc**TTG**
yhxC	144004->144861	285	30.8	7.42	GAAAaGgaGTGATttca**TTG**
comK	145415->145993	192	22.4	7.77	AGgAtGGAGGccATaat**ATG**
yhxD	146040<-146939	299	31.9	4.64	AAAGGAGcgttgCtg**ATG**
yhjA	147156->147425	89	9.8	9.99	GagaGTGAatcgtc**ATG**
yhjB	147468<-148937	489	52.8	9.42	AaAAAGGAGGaagcaga**ATG**
yhjC	148934<-149134	66	7.4	7.14	AAGGAGGattct**ATG**
yhjD	149342<-149704	120	14.5	5.87	AGAAAGaAGGaGtcaat**ATG**
yhjE	149857->150480	207	23.3	10.12	GtAAGGAGtatAa**ATG**
yhjF	150482->150988	168	19.0	9.72	AtgAtGGAGGgagaCagtaac**ATG**
yhjG	151171->152667	498	54.2	6.96	AAAGGAGtgGtga**atg**ATG
yhjH	152744->153271	175	20.4	7.55	AaAAAGGAtGgaAaacccg**ATG**
yhjI	153429<-154634	401	44.9	8.70	AtaGgGGTGtaatga**ATG**
yhjJ	154706<-155758	350	39.3	6.50	AGGAGGaaATaaaa**ATG**
yhjK	155761<-156621	286	33.2	5.52	AAtGGAGGgacTgtttc**ATG**
yhjL	156593<-157918	441	50.1	6.24	AttGGAGGTacTgttc**ATG**
yhjM	158022->159011	329	37.7	6.91	AAGGAaGgGAaaat**ATG**
yhjN	159225<-160379	384	41.0	9.64	AGgAAGaAGGgttTtaca**TTG**
yhjO	160486<-161691	401	44.1	9.53	GAAAGGcGGcGATCac**ATG**
yhjP	161805->163532	575	66.4	6.40	GtcgGGAGGTGcgggga**TTG**
yhjQ	163562<-163888	108	11.8	5.03	AtAcAGGgGGaatcaacc**ATG**
yhjR	164006<-164443	145	17.2	6.21	AGAAtGGAGtTGAatcccc**TTG**
addB	164627->168127	1166	134.6	5.56	AagAgaGgGGTctTCtaat**TTG**
addA	168114->171812	1232	141.1	5.26	AaAAAGGAGGcGgatggca**ATG**

For legend see next page

Legend to Table II.2. In the column "ORF", bold letters represent genes which were already characterized from other studies. In the column "endpoints", a right-pointing arrow means that the ORF is transcribed clockwise on the chromosome; left-pointing arrows indicate putative genes that are transcribed counterclockwise. In the column "S-D consensus sequence and initiation codon", bases that are complementary to the 16S rRNA are indicated with capitals; the putative initiation codon is indicated in bold capitals. N.P. = Not present. When an alternative possible initiation codon was found, it is also indicated in bold.

Updating and correction of the genetic map of the *prkA-addAB* region

From our cloning and sequencing data, it became clear that the genetic map of this region (Anagnostopoulos *et al.*, 1993) contained several errors. The corrected genetic/physical map of the region is presented in Fig. II.2. The corrected positions of genes are presented in degrees relative to the origin of replication. We calculated the size of a DNA fragment corresponding to one degree on the chromosome by dividing the determined genome size of 4,214,807 bp (Kunst *et al.*, 1997) by 360. According to this calculation, one degree on the chromosome corresponds to 11,708 bp.

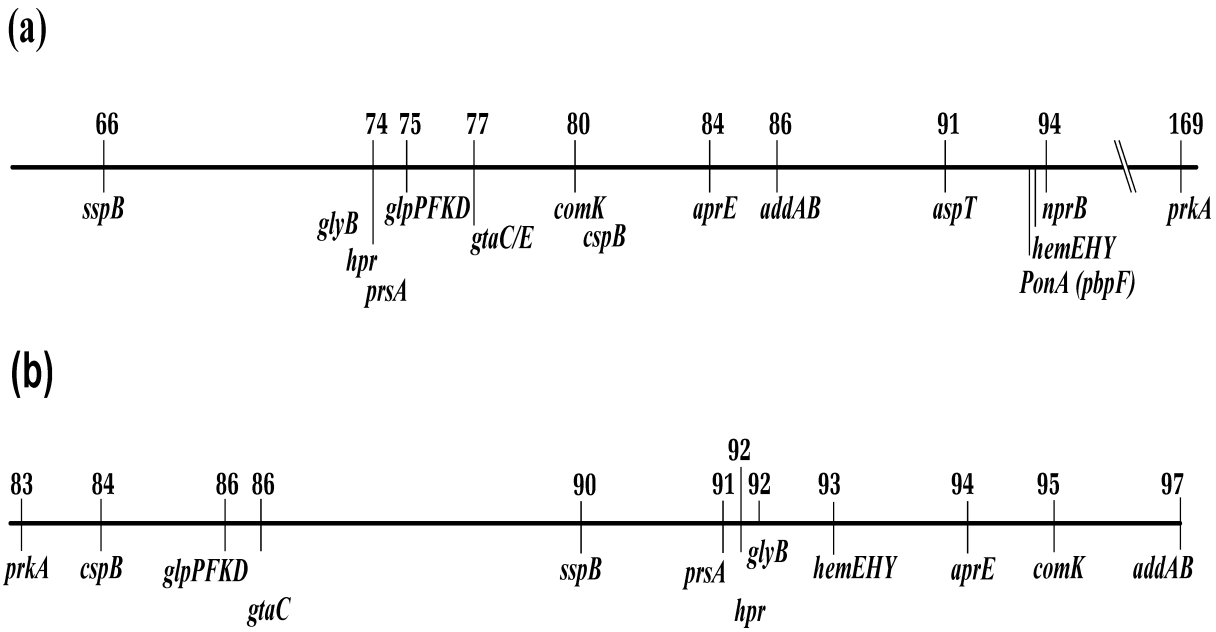


Fig. II.2. Update of the genetic map of the *prkA* to *addAB* region. (a): Part of the genetic map of the *B. subtilis* chromosome according to Anagnostopoulos *et al.* (1993). Numbers above the line representing the map indicate the position, in degrees, relative to the origin of replication. (b): Corrected map of the region based on sequence data. Numbers above the line indicate positions in degrees relative to the origin of replication, as deduced from the total genome sequence, with one degree calculated to be 11,708 bp.

Deduced gene products and similarity analysis

All deduced amino acid sequences from putative genes within this region were compared to known protein sequences in public databanks, and to the putative proteins encoded by the *B. subtilis* chromosome.

The similarity of deduced protein products from the sequenced region with known protein sequences in the databanks is presented in Table II.3. On the basis of similarity to known proteins, we propose that *yhxB* corresponds to the *gtaC* marker and *yhaF* to the *glyB* marker (see also below).

We classified all ORFs according to their putative function (the results of which were already summarised in Fig. II.1). The different global classes of functions are mainly as described by (Kunst *et al.*, 1997). Cell envelope and cellular processes include: proteins involved in cell wall metabolism, transport/binding proteins, lipoproteins, and proteins involved in membrane bioenergetics, mobility, chemotaxis, secretion and sporulation. 'Intermediary metabolism' includes: proteins involved in the metabolism of carbohydrates, amino acids, nucleotides and nucleic acids, and coenzymes and prosthetic groups. 'Information pathways' includes: proteins involved in DNA synthesis, restriction/modification, recombination and repair, RNA synthesis, and protein synthesis. 'Other' includes: functions like antibiotic production, drug (-analog) sensitivity, and adaptation to atypical conditions ("stress proteins").

Table II.3. Deduced ORF products, the number of paralogous sequences, and their similarities with protein sequences in public databases

ORF product	**# PL**	**Similar protein(s) in databases**	**Database Accession number**	**% Identity, S.-W. score, # aa overlap**
YhbE	2	=YzdA from *Bacillus subtilis*	SP|P39132	100
YhbF	1	=YzdB from *B. subtilis*	SP|P39133	100
PrkA	0	=Protein kinase A PrkA from *B. subtilis*	SP|P39134	100
YhbH	0	=YzdC from *B. subtilis* and	SP|P45742	100
		Hypo YzdC from *Escherichia coli*	D90822/g1736412	27, 523, 411
YhbI	4	Multiple antibiotic resistance operon regulatory protein MarR from *Salmonella typhimurium*	U54468/g1293698	30, 189, 138
YhbJ	0	Multidrug resistance protein A (EmrA) from *E. coli*	SP|P27303	29, 121, 75
YzdF	1	Multidrug resistance protein A (EmrA) from *E. coli*	SP|P27303	31, 216, 114
YhcA	5	Multidrug resistance protein B (EmrB) from *E. coli*	SP|P27304	29, 732, 431
YhcB	1	Trp repressor-binding protein WrbA from *E. coli* and	SP| P304849	32, 294, 189
		flavodoxin from *Clostridium acetobutylicum*	SP| P18855	31, 210, 119
YhcC	3	None		
YhcD	1	None		
YhcE	0	None		
YhcF	5	GntR regulator family, like KorA from *Streptomyces lividans*	SP| P22405	28, 161, 88
		and FarA from *E. coli*. (YhcF is much shorter, spanning only the N-terminal half of these proteins)	SP| P13669	39, 156, 71
YhcG	53	ABC transporters: CysA from *Synechococcus* sp. and	SP| P14788	34, 369, 212
		NosF from *Pseudomonas stutzeri*	SP| P19844	31, 373, 222
YhcH	29	ABC transporters: NosF from *P. stutzeri*, BcrA from *Bacillus licheniformis*, StpC (*Staphylococcus aureus*) and	SP| P19844	34, 544, 307
			SP| P42332	37, 683, 303
		YhcG, the preceding ORF on the *B. subtilis* chromosome	E|Z30588/g459256	37, 535, 226
YhcI	1	Membrane protein NosY from *P. stutzeri*,	SP| P19845	25, 123, 231
		BcrB from *B. licheniformis*, and	SP|P42333	24, 139, 183
		SmpC from *Staphylococcus aureus*	E| Z30588/g459257	26, 242, 219
CspB	4	Cold shock protein B	U58859/g1336658	100

YhcJ	1	Lipoprotein-28 precursor NlpA from *E. coli*	SP| P04846	30, 374,	257
YhcK	0	Hypothetical proteins from *Streptomyces ambofaciens* and *Vibrio anguillarum* (ORF3)	SP| P36892 U17054/g576657	29, 166, 34, 313,	162 201
YhcL	0	Proton/sodium-glutamate symport protein GltT from *Bacillus caldotenax*	SP| P24944	27, 483,	421
YhcM	0	None			
YhcN	0	CS3 pili biogenesis protein from *E. coli*	SP| P15487	22, 81,	98
YhcO	3	None			
YhcP	2	None			
YhcQ	0	Spore coat protein F (CotF) from *B. subtilis*, mainly in the C-terminal half	SP| P23261	23, 122,	90
YhcR	0	The C-terminal half: UDP-sugar hydrolase precursor UshA from *E. coli*, and 5'-nucleotidase precursor from *Bos taurus* (bovine)	SP| P07024 SP| Q05927	28, 490, 22, 383,	572 546
YhcS	0	None			
YhcT	1	DRAP deaminase from *Saccharomyces cerevisiae* and a family of hypothetical proteins of which YceC from *E. coli* is also a member	PIR| S50972 SP| P33643	24, 274, 39, 529,	246 254
YhcU	0	None			
YhcV	9	IMP dehydrogenase GuaB from *B. subtilis* and AcuB (involved in acetoin utilization) from *B. subtilis*	SP| P21879 SP| P39066	31, 193, 27, 160,	118 121
YhcW	3	Phosphoglycolate phosphatase from *Alcaligenes eutrophus* and a family of hypothetical proteins (like YieH from *E. coli*)	SP| P40852 SP| P31467	25, 179, 27, 204,	186 181
YhcX	0	Nitrilase 2 from *Arabidopsis thaliana* and a hypothetical protein from *S. cerevisiae*	SP| P32962 PIR| S51459	34, 156, 27, 326,	103 292
YhxA	6	DAPA aminotransferase (BioA) from *Bacillus sphaericus*	SP| P22805	34, 839,	446
GlpP	1	=Glycerol operon regulator GlpP from *B. subtilis*	SP|P30300	100	
GlpF	2	=Glycerol uptake facilitator GlpF from *B. subtilis*	SP|P18156	100	
GlpK	2	=Glycerol kinase GlpK from *B. subtilis*	SP|P18157	100	
GlpD	0	=Glycerol-3-phosphate dehydrogenase GlpD from *B. subtilis*	SP|P18158	100	
YhxB	1	Phosphomannomutase or phosphoglucomutase from *Mycoplasma pirum* (and many other organisms)	PIR|E53312	28, 793,	564
YhcY	0	Sensory transduction kinase DegS from *B. subtilis*	SP|P13799	31, 261,	221
YhcZ	15	Transcriptional regulator DegU from *B. subtilis*	SP|P13800	39, 517,	219
YhdA	2	Hypo YieF from *E. coli*	SP|P31465	26, 174,	136
YhdB	0	None			
YhdC	0	None			
YhdD	2	Phosphatase-associated protein PapQ from *B. subtilis*	GB|U38819	50, 943,	316
YhdE	1	Hypo YjeB from *E. coli*	SP|P40610	44, 393,	142
YgxB	0	=YgxB from *B. subtilis* (partial) Hypo from *Synechococcus* sp.	SP|P37874 PIR|S20924	100 28, 248,	 173
SpoVR	0	=Stage V sporulation, SpoVR from *B. subtilis*	SP|P37875	100	
PhoAIV	1	=Alkaline phosphatase, PhoAIV from *B. subtilis*	SP|P19406	100	
PapQ	4	=Phosphatase-associated protein, PapQ from *B. subtilis*	EMB|U38819	100	
CitR	2	=Negative regulator for *citA*, CitR from *B. subtilis*	SP|P39127	100	
CitA	2	=Citrate synthase I, CitA from *B. subtilis*	SP|P39119	100	
YhdF	21	Glucose and ribitol dehydrogenase from barley	GP|S7226	52, 952,	286
YhdG	5	Hypo from *Mycobacterium tuberculosum* and Cationic amino acid transporter from *Homo sapiens*	Z79702/g264157; D29990/g849051	41, 1269, 36, 893,	464 435
YhdH	1	Hypo YG90 from *Haemophilus influenzae*	SP|P455320	36, 1064,	457
YhdI	5	Probable rhizopine catabolism regulatory protein MocR from *Rhizobium meliloti*, and aminotransferase from *Sulfolobus solfataricus*	SP|P49309 E283830/g17 07790	34, 897, 27, 397,	481 370
YhdJ	0	Regulator of alkylphosphate uptake PhnO from *E. coli*	SP|P16691	34, 136,	82
YhdK	4	None			
YhdL	0	None			

YhdM	7	Putative RNA polymerase sigma factor YbbL from *B. subtilis*	D84214/g1256141	31, 280, 160
YhdN	5	Hypo YxbF from *B. subtilis*	SP|P46336	37, 704, 311
		Potassium channel ß2 subunit from *Homo sapiens* (human)	U33429/g995761	30, 402, 334
YhdO	0	Hypo from *Synechocystis* sp.	D90915/g1653690	26, 200, 180
YhdP	4	YhdT (this paper) from *B. subtilis*	this paper	61, 1687, 430
		Hemolysin from *Synechocystis* sp.	D90914/g1653594	30, 677, 441
YhdQ	2	Hypo HI1623 from *H. influenzae*	SP|P45277	33, 184, 120
		Mercury resistance regulatory protein MerR from *Thiobacillus ferrooxidans*	SP|P22896	35, 154, 87
YhdR	4	Aspartate aminotransferase from *Methanococcus jannaschii*	U67459/g1592252	30, 520, 391
YhdS	0	Hypo from Fowlpox virus (small internal fragment)	SP|P21973	44, 63, 25
YhdT	4	YhdP (this paper) from *B. subtilis*	this paper	61, 1687, 430
		Hemolysin from *Synechocystis* sp.	D90914/g1653594	31, 683, 439
YhdU	2	NADH-plastoquinone oxidoreductase chain 2 (chloroplast) from *Marchantia polymorpha*	SP|P06257	24, 125, 122
YhdV	5	None		
YhdW	2	Glycerol diester phosphodiesterase (GlpQ) from *B. subtilis*	SP|P37965	38, 575, 252
YhdX	0	Hypo Human transposon L1.1 ORF1	M80340/g339770	32, 60, 34
YhdY	0	Hypo MJ1143 from *M. jannaschii*	g1591775	27, 550, 357
YhdZ	0	Lac repressor LacR from *S. aureus*	M32103/g845686	36, 446, 251
YheN	1	Hypo Yfu2 from *B. stearothermophilus*	SP|Q04729	32, 305, 205
YheM	2	D-amino acid aminotransferase from *B. licheniformis*	U26947/g857561	64, 1179, 275
YheL	1	Na(+)/H(+) antiporter from *B. firmus*	SP|P27611	53, 1377, 390
YheK	1	Hypo YxiE from *B. subtilis*	SP|P42297	30, 230, 166
YheJ	0	None		
YheI	11	Multidrug resistance-like ATP binding protein MDL from *E. coli*	SP|P30751	37, 1134, 507
YheH	9	Multidrug resistance-like ATP binding protein MDL from *E. coli*	SP|P30751	40, 1341, 519
YheG	2	Flavin reductase FLR from *Bos taurus* (Bovine)	SP|P52556	27, 211, 208
YheF	0	None		
SspB	3	=Small, acid-soluble spore protein B, SspB from *B. subtilis*		100
YheE	1	None		
YheD	0	None		
YheC	0	Central part of hypo MJ0776 from *M. jannaschii*	U67522/g1499596	32, 142, 123
YheB	0	Hypo orf sll0412 from *Synechocystis* sp.	D64001/g1001108	26, 335, 405
YheA	3	None		
YhaZ	0	None		
YhaY	1	None		
YhaX	2	Hypo YcsE from *B. subtilis*	SP|P42962	27, 266, 257
		Hypo Cof protein from *E. coli*	SP|P46891	26, 234, 251
YhaW	1	None		
YhaV	1	Anaerobic coproporphyrinogen III oxidase HemN from *H. influenzae*. (see also text)	SP|P43899	27, 404, 332
YhaU	1	Na(+)/H(+) antiporter from *Enterococcus hirae*	SP|P26235	26, 410, 386
YhaT	2	C-terminal part of hypo form *Synechocystis* sp.	D64006/g1001375	29, 138, 84
YhaS	0	None		

YhaR	4	Enoyl-CoA-hydratase from *Rhodobacter capsulatus*	SP|P24162	33, 390,	246
YhaQ	24	ATP-binding transport proteins (ABC-transporter) from *B. firmus* (hypothetical) and from *M. jannaschii*	SP|P26946 U67545/g1499865	62, 1168, 42, 690,	266 260
YhaP	0	N-terminal part to methylmalonyl-CoA mutase homolog, MutX from *B. firmus* and to *M. jannaschii* hypo MJ1024 (full length)	SP|P26947 U67545/g1499866	45, 168, 25, 403,	56 402
YhaO	0	Hypo sll0021 from *Synechocystis* sp. Hypo MJ1323 from *M. jannaschii* SbcD from *E. coli* SbcD homolog from *B. subtilis*	D64000/g100 1554; U67572 /g1591963 SP|P13457 SP|P23479	26, 228, 25, 243, 25, 154, 24, 141,	310 306 276 277
YhaN	0	Hypo Orf X from *S. aureus* (from aa 600 of YhaN) Exonuclease subunit SbcC from *E. coli* Rad50 of multiprotein complex implicated in recombinational DNA repair from *H. sapiens*	U21636/g710 421; SP|P134 58; U63139/ g1518806	25, 415, 20, 313, 21, 234,	358 856 821
YhaM	0	Cmp-binding factor 1 from *S. aureus* Hypo MJ0837 from *M. jannaschii*	U21636/g710 422; U67528/ g1499663	52, 1137, 32, 196,	300 144
YhaL	1	None			
PrsA	4	=Protein export protein PrsA from *B. subtilis*	SP|P24327	100	
YhaK	1	None			
YhaJ	2	None			
YhaI	0	None			
Hpr	0	=Protease production regulatory protein Hpr from *B. subtilis*	SP|P11065		
YhaH	2	Clone pSJ7 product from *B. subtilis* (from aa 57 of yhaH) Hypo YtxH from *B. subtilis* Apolipoprotein A-I (Apo-AI) precursor from *Oryctolagus cuniculus* (Rabit)	S70232/g547 157 SP|P40780 SP|P09809	79, 229, 25, 178, 29, 128,	42 113 107
YhaG	1	Glycine Betaine/L-proline transport system permease protein ProW from *E. coli* (only C-terminal half; see also text)	SP|P14176	20, 86,	148
YhaF	0	Phosphoserine aminotransferases from *B. circulans*, *Spinacia oleracea* (SerC), *A. thaliana*, *H. influenzae* (SerC), Rabit (SerC), and *E. coli* (SerC),	gnl|PID|e123178 SP|P52877 D88541/g1665831 SP|P44336 SP|P10658 SP|P23721	54, 1329, 50, 1162, 50, 1156, 46, 985, 44, 1000, 44, 953,	357 363 362 360 362 364
YhaE	1	Member of the HIT family of proteins, with members from *M. jannaschii*, *Mycoplasma pneumoniae*, *Borrelia burgdorferi*, *Mycoplasma genitalium*, and *S. solfataricus*	U67530/g1499694 /g1674261 U49938/g1753229 SP|P47378 Y08256/g1707769	50, 372, 49, 365, 50, 354, 46, 352, 42, 300,	128 110 113 134 105
EcsA	60	=ABC-type transporter ATP-binding protein EcsA from *B. subtilis*	SP|P55339	100	
EcsB	0	=Hypothetical integral membrane protein EcsB from *B. subtilis*	SP|P55340	100	
EcsC	1	=Protein EcsC from *B. subtilis*	SP|P55341	100	
YhaA	4	N-acyl-L-amino acid amidohydrolase from *B. stearothermophilus*	SP|P37112	43, 864,	305
YhfA	0	Anaerobic carrier for dicarboxylates, DcuC from *E. coli*	X99112/g252616	24, 194,	476
YixB	0	=Hypo YixB from *B. subtilis* (fragment)	SP|P38048	100,	67
YixC	1	=Hypo YixC from *B. subtilis*	SP|P38049	100	
PbpF	3	=Penicillin-binding protein PbpF from *B. subtilis*	SP|P38050	100	
HemE	0	=Uroporphyrinogen decarboxylase HemE (=DcuP) from *B.*	SP|P32395	100	

		subtilis			
HemH	0	=Ferrochelatase HemH from *B. subtilis*	SP|P32396	100	
HemY	0	=Coproporphyrinogen III oxidase HemY from *B. subtilis*	SP|P32397	100	
YixD	5	=Hypo YixD from *B. subtilis*	SP|P32398	100	
YixE	0	=Hypoth. protein in HemY 3'region (orfB; fragment) from *B. subtilis*, and	SP|P32399 SP|P49022	100, 23, 742,	145 885
		phage infection protein from *Lactococcus lactis*			
YhfB	1	Beta-ketoacyl-acyl carrier protein (FabH) from *E. coli*,	SP|P24249	39, 741,	319
		Porphyra purpurea, and others	SP|P51196	36, 720,	323
YhfC	1	None			
YhfD	1	Part of metallothionein isoform Ia from *Callinectes sapidus*	g1176448	29, 63	31
YhfE	1	Endoglucanase CelM from *Clostridium thermocellum*	g1097207	26, 304,	345
YhfF	1	Late embryogenesis abundant protein group 3 from *Tritium aestivum* (wheat); partial	PIR|S33616	29, 99,	96
YhfG	2	Proton/sodium-glutamate symport protein from *B. stearothermophilus* (GltT),	SP|P24943 SP|P24944	64, 1489, 63, 1478,	344 344
		B. caldolyticus (GltT),	SP|P21345	57, 1272,	341
		E. coli (GltP), and	SP|P39817	46, 1037,	349
		B. subtilis (GltP)			
YhfH	0	Small toxin SCXI from *Mesobuthus tamulus sindicus* (scorpion), and low similarity to many Zn-finger proteins. This orf contains the Zinc-finger motif CXXC…CXXC	SP|P15229	52, 71,	23
YhfI	1	Arylsulfatase precursor from *Mycobacterium leprae*	U00014/g466916	29, 337,	249
YhfJ	0	Lipoate protein ligase from *M. pneumoniae* (LplA),	U00089/g167	34, 758,	327
		M. genitalium (LplA), and from	4137	34, 700,	336
		E. coli (LplA)	SP|P47512 SP|P32099	35, 596,	315
YhfK	3	Hypo YM9582.15 from *S. cerevisiae*	PIR|S54466	38, 462,	225
YhfL	19	Long-chain-fatty-acid CoA ligase LcfA from *E. coli*,	SP|P29212	40, 1173,	533
		H. influenzae (LcfA)	SP|P46450	36, 1040,	532
YhfM	0	None			
YhfN	0	Hypo YzoA from *B. subtilis* (=fragment of YhfN),	SP|P40769	100,	42
		Hypo YJ87 from *S. cerevisiae*	SP|P47154	25, 382,	419
AprE	5	=Subtilisin (extracellular alkaline serine protease) from *B. subtilis*	SP|P04189	100	
YhfO	3	Hypo Y677 from *H. influenzae*	SP|P44036	32, 234,	135
YhfP	3	Hypo YhdH from *E. coli*	SP|P26646	47, 976,	325
YhfQ	7	Iron(III)dicitrate transport protein from *E. coli* (FecB), and	PIR|S56515	32, 486,	282
		from	D90899/g165	28, 434,	328
		Synechocystis sp.	1665		
YhfR	0	Hypo o215b from *E.coli*,	PIR|S56619	32, 307,	189
		Probable phosphoglycerate mutase (Pgm) from *E.coli*, and	SP|P36942	32, 303,	189
		Pgm from *Treponema pallidum*	U55214/g1777938	38, 221,	100
YhfS	2	Acetyl-CoA—acetyltransferase ThiL from *Thiocystis violacea*,	SP|P45363	39, 790,	392
		Chromatium vinosum,	SP|P45369	38, 788,	394
		Alcaligenes eutrophus and	SP|P14611	40, 773,	392
		B. subtilis	SP|P45855	38, 729,	391
YhfT	8	Long-chain-acyl—CoA synthetase from *B. subtilis*,	Z75208/g177	29, 590,	539
		Bile acid-CoA ligase from *Eubacterium* sp.,	0038	28, 546,	487
		Long-chain-fatty-acid-CoA ligase (LcfA) from *E. coli*	SP|P19409 SP|P29212	26, 455,	479
YhfU	4	BioY (biotin synthesis) from *B. sphaericus*	SP|P22819	31, 250,	186
YhfV	0	Methyl-accepting chemotaxis protein from *Halobacterium salinarium* (HtB),	U75436/g1654420	26, 496, 30, 383,	454 288
		B. subtilis (TlpC),	SP|P39209	30, 377,	289
		B. subtilis (TlpB), and from	SP|P39217	30, 366,	250
		B. subtilis (TlpA)	SP|P39216		

YhfW	0	Oxidoreductase OrdL from *E. coli*	U38543/g1054921	20, 308,	431
YhxC	22	=YhxC from *B. subtilis* (fragment)	SP|P40397	100,	114
		Glucose and ribitol dehydrogenase homolog from *Hordeum vulgare* (barley)	GB|S72926	56, 1002,	295
ComK	0	=Competence protein K from *B. subtilis*	SP|P40396	100	
YhxD	23	=YhxD from *B. subtilis* (fragment)	SP|P40398	100,	140
		Hypo ORF_o294 *E. coli*	U26377/g882532	64, 1281,	292
		Glucose and ribitol dehydrogenase homolog from *H. vulgare* (barley)	GB|S72926	43, 691,	288
YhjA	3	None			
YhjB	1	Proline permease PutP from *S. typhimurium*	GB|S72926	25, 400,	495
YhjC	1	None			
YhjD	1	None			
YhjE	0	Hypo YqeD from *B. subtilis*	D84432/g1303784	22, 225,	190
YhjF	4	Type I signal peptidase from *B. caldolyticus* (SipC)	SP|P41027	50, 497,	159
		and from *B. subtilis* (SipT)	U45883/g1518930	42, 394,	161
YhjG	0	Tetracycline 6-hydroxylase from *Streptomyces aureofaciens*	PIR|JC4098	40, 1080,	493
		and pentachlorophenol 4-momooxi-genase from *Flavobacterium* sp.	SP|P42535	32, 893,	476
YhjH	1	Hypo YzhA from *B. subtilis* and	SP|P40762	42, 362,	143
		multidrug resistance operon repressor MexR from *Pseudomonas aeruginosa*	U23763/g886021	24, 103,	71
YhjI	0	Hypo YOL173w from *S. cerevisiae* and	EMBL|Z74879	25, 315,	375
		glucose and galactose transporter from *Brucella abortus*	22, 227,	365	
			U43785/g1171339		
YhjJ	2	Myo-inositol 2-dehydrogenase MI2D from *B.subtilis*	SP|P26935	26, 237,	262
		and glucose-fructose oxidoreductase Gfo from *Zymomonas mobilis*	Z80356/g1657416	23, 200,	307
YhjK	0	Hypo YpdA from *B. strearothermophylus*	SP|P21878	37, 173,	82
		and phosphoserine phosphatase SerB from *H. influenzae*	SP|P44997	23, 102,	230
YhjL	1	Pleiotropic regulatory protein DegT from *B. stearothermophilus* and spore coat polysaccharide biosynthesis protein SpsC from *B. subtilis*	SP|P15263	37, 695,	369
			SP|P39623	33, 676,	392
YhjM	10	Transcriptional repressor CytR from *E. coli*,	SP|P06964	33, 609,	330
		degradation activator DegA from *B. subtilis*, and	SP|P37947	31, 568,	331
		catabolite control protein CcpA from *B. subtilis*	SP|P25144	30, 581,	332
YhjN	0	Hypo f363 from *E. coli* and	gi1786933	27, 333,	297
		proton antiporter efflux protein from *Mycobacterium smegmatis*	U40487/g1110518	23, 95,	271
YhjO	1	Hypo YqjV from *B. subtilis* and	D84432/g1303973	23, 423,	392
		multidrug resistance protein 1 (BMR1) and		25, 307,	381
		multidrug resistance protein 2 (BMR2) from *B. subtilis*	SP|P33449	24, 274,	385
			SP|P39843		
YhjP	0	Hypo YabN from *E. coli* and	SP|P33595	25, 551,	586
		oligopeptide-binding protein AppA from *B. subtilis*	SP|P42061	26, 223,	298
YhjQ	1	Polyferredoxin from *M. jannaschii*	U67560/g1591821	24, 115,	78
YhjR	0	Nigerythrin from *Desulfovibrio vulgaris*	U71215/g1616801	25, 112,	128
AddB	0	=ATP-dependent deoxyribonuclease subunit B from *B. subtilis*	SP|P23477	100	
AddA	0	=ATP-dependent deoxyribonuclease subunit A from *B. subtilis*	SP|P23478	100	

Proteins that were previously known are indicated in bold. Indicated are the percentage identity, the Smith-Waterman score (S.-W score), and the length of the homologous region in amino acids. # PL: the number of paralogous sequences within the *B. subtilis* genome. Hypo = hypothetical protein (no experimental evidence for its function). SP = Swiss Prot; GB = GenBank; E = EMBL; GP = GenPept.

References

Anagnostopoulos, C., Piggot, P. J. & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. *In Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, pp. 425-461. Edited by A. L. Sonenshein, J. A. Hoch, and R. Losick. Washington, DC, American Society for Microbiology.

Barnes, W. M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from **Lambda** bacteriophage templates. *Proc.Natl.Acad.Sci.USA* **91**, 2216-2220.

Beall, B. & Moran Jr, C. P. (1994). Cloning and characterization of *spoVR*, a gene from *Bacillus subtilis* involved in spore cortex formation. *J.Bacteriol.* **176**,

Beijer, L., Nilsson, R.-P., Holmberg, C., & Rutberg, L. (1993). The *glpP* and *glpF* genes of the glycerol regulon in *Bacillus subtilis*. *J.Gen.Microbiol.* **139** , 349-359.

Biaudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S. D., & Bessières, P. (1996). Computerized genetic map of *Bacillus subtilis*. *Microbiol.* **142**, 2669-2729.

Bron, S. (1990). Plasmids. *In* molecular biological methods for *Bacillus*, pp. 75-174. Edited by C. R. Harwood and S. M. Cutting. Chichester, John Wiley & Sons.

Bron, S. & Venema, G. (1972). Ultraviolet inactivation and excision repair in *Bacillus subtilis*. I. Construction and characterization of a eightfold auxotrophic strain and two ultraviolet-sensitive derivatives. *Mutat.Res.* **15**, 1-10.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., & Venter, J. C. (1996). Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073.

Cheng, S., Chang, S. Y., Gravitt, P., & Respess, R. (1994). Long PCR. *Nature* **369**, 684-685.

Dear, S. & Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19**, 3907-3911.

Fischer, C., Geourjon, C., Bourson, C., & Deutscher, J. (1996). Cloning and characterization of the *Bacillus subtilis prkA* gene encoding a novel serine protein kinase. *Gene* **168**, 55-60.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A., & Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**, 546-567.

Hansson, M. & Hederstedt, L. (1992). Cloning and characterization of the *Bacillus subtilis hemEHY* gene cluster, which encodes protoheme IX biosynthetic enzymes. *J.Bacteriol.* **174**, 8081-8093.

Harford, N., Lepesant-Kejzlarova, J., Lepesant, J.-A., Hamers, R., and Dedonder, R. (1976). Genetic circularity and mapping of the replication origin region of the *Bacillus subtilis* chromosome. *In* Microbiology, pp. 28-34. Edited by D. Schlesinger. Washington, D.C., American Society for Microbiology.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420-4449.

Holmberg, C., Beijer, L., Rutberg, B., & Rutberg, L. (1990). Glycerol catabolism in *Bacillus subtilis*: nucleotide sequence of the genes encoding glycerol kinase (*glpK*) and glycerol-3-phosphate dehydrogenase (*glpD*). *J.Gen.Microbiol.* **136**, 2367-2375.

Hulett, F. M., Kim, E. E., Bookstein, C., Kapp, N. V., Edwards, C. W., & Wyckoff, H. W. (1991). *Bacillus subtilis* alkaline phosphatase III and IV. Cloning, sequencing, and comparisons of deduced amino acid sequence with *Escherichia coli* alkaline phosphatase three-dimensional structure. *J.Biol.Chem.* **266**, 1077-1084.

Ish-Horowicz, D. & Burke, F. J. (1981). Rapid and efficient cosmid cloning. *Nucleic Acids Res.* **9**, 2989-2999.

Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J.Mol.Biol.* **220**, 631-648.

Jin, S. & Sonenshein, A. L. (1994). Identification of two distinct *Bacillus subtilis* citrate synthase genes. *J.Bacteriol.* **176**, 4669-4679.

Kontinen, V. P., Saris, P., & Sarvas, M. (1991). A gene (*prsA*) of *Bacillus subtilis* involved in a novel late stage of protein export. *Mol.Microbiol.* **5**, 1273-1283.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Düsterhöft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, H., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., & Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Mandel, M. & Higa, A. (1970). Calcium-dependent bacteriophage DNA infection. *J.Mol.Biol.* **53**, 159-162.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F., & Zollner, A. (1997). Overview of the yeast genome. *Nature* **387**, 7-9.

Moszer, I., Glaser, P., & Danchin, A. (1995). SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiol.* **141**, 261-268.

Noback, M. A., Terpstra, P., Holsappel, S., Venema, G., & Bron, S. (1996). A 22 kb DNA sequence in the *cspB-glpPFKD* region at 75° on the *Bacillus subtilis* chromosome. *Microbiol.* **142**,

O'Brien, C. (1997). Entire *E. coli* genome sequenced -- at last. *Nature* **385**, 472-472.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc.Natl.Acad.Sci.USA* **85**, 2444-2448.

Perego, M. & Hoch, J. A. (1988). Sequence analysis and regulation of the *hpr* locus, a regulatory gene for protease production and sporulation in Bacillus subtilis. *J.Bacteriol.* **170**, 2560-2567.

Sambrook, J., Fritsch, E. F., and Maniatis T. (1989). Molecular cloning: a laboratory manual. 2.. Cold Spring Harbor, Cold Spring Harbor Laboratory.

Sanger, F., Nicklen, S., & Coulsen, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.USA* **74**, 5467

Stahl, M. L. & Ferrari, E. (1984). Replacement of the *Bacillus subtilis* subtilisin structural gene with an in vitro-derived deletion mutation. *J.Bacteriol.* **158**, 411-418.

CHAPTER III

In silico analysis of the 172 kb DNA region from 83° to 97° of the *Bacillus subtilis* chromosome: protein localisation, paralogs and dysfunctional genes

III.1. Summary

In this chapter, an overview is presented of the logical continuation of the *in silico* analysis of the DNA fragment, described in chapter II, in which the genes have been annotated, and their protein sequences were deduced. In this chapter, we focus on several aspects. First, the localisation of proteins encoded by genes from the sequenced region is predicted. The putative proteins were analysed for the presence of transmembrane helices, lipomodification signals, and signal sequences typical of secreted proteins. Second, the availability of the entire *B. subtilis* genome sequence enabled us to compare ORFs in the *prkA* to *addAB* region with all coding sequences in the entire genome. This revealed the existence of numerous paralogs to ORFs in this region. About two thirds of the putative genes in the *prkA* to *addAB* region seem to have at least one paralog in the *B. subtilis* genome. Closer investigation of paralogs revealed that - on the secondary structure level of the deduced proteins - modules of protein architecture could be detected. This is exemplified with six ABC-type ATP-binding transporter proteins. Also, several examples are presented of ORFs from the region that are most likely dysfunctional.

III.2. Introduction

The aim of this chapter is to show that sequencing, annotation and homology analysis of a DNA region only reveals a fraction of the available information that can be extracted from such data. This is particularly the case when the entire genome of the organism from which the DNA region has been sequenced, is known. When a sequence is annotated, the DNA can be analysed for additional features like expression signals, such as promoters and regulator binding sequences, but also topics like the evolutionary origin of the chromosome and other phylogenetic questions can be addressed. The putative proteins can be further analysed by, for instance, domain searches (such as ATP binding domains, cofactor binding domains), paralog analysis, search for localisation signals, such as signal sequences, transmembrane helices and lipomodification signals. Besides mere homology searches, all these analyses together can yield valuable information about the possible function of an unknown gene and give insight on possible (experimental) ways how to uncover that function. In this chapter, a number of the

above-mentioned topics will be addressed. These concern analysis of localisation signals (e.g. transmembrane helices, signal sequences, and lipomodification signals) and paralog frequencies. Paralogs are defined as homologous proteins resulting from gene duplications within an organism, while orthologs are genes, sharing a common ancestor, that perform the same role in different species (Henikoff *et al.*, 1997). One of the most striking findings that arose from recent genome analyses is the abundance of paralogous relationships between genes. Some attention will also be given to the presence of probable dysfunctional genes in this region. Unfortunately, it is still not possible to use computer-assisted searches for possible *B. subtilis* promoters, since the promoters are too diverse - *B. subtilis* has at least fourteen ORFs encoding RNA polymerase sigma factors - and too ambiguous in sequence and spacing of the –35 and –10 sequences to allow their identification by simple search-strings. With *E. coli* sequences, promoter searches are possible with the aid of a neural network trained for the *E. coli* sigma A-dependent consensus promoter.

III.3. Methods

Sequence comparisons were executed using the FASTA program of Pearson and Lipman (1988). Multiple sequence alignments were done using the ClustalW program at the www site of EBI (http://www2.ebi.ac.uk/clustalw/). Signal peptide predictions were done using the SignalP program at the web server of the Center for Biological Sequence Analysis of the Technical University of Denmark (at http://www.cbs.dtu.dk/), which can predict signal peptides and their cleavage sites in gram-positive, gram-negative, and eukaryotic amino acid sequences (Nielsen *et al.*, 1997). Prediction of putative membrane localisation and topology were done using the TopPred 2 program (at web address: http://www.biokemi.su.se/~server/toppred2/) of the Theoretical Chemistry group of Stockholm University (von Heijne, 1992).

III.4. Results and Conclusions

Protein localisation

In order to assess where the (putative) proteins encoded on the *prkA-addAB* region will be localised, we searched the deduced amino acid sequences of genes from this region for signatures indicating lipomodification, secretion and/or membrane association.

In Fig.III.1, the frequency distribution of proteins with membrane-spanning domains from this region is shown. This distribution also includes proteins with a signal sequence, since the Toppred2 program does not discriminate between these two protein varieties. According to the Toppred2 predictions, a surprisingly high fraction (51%) of the deduced proteins from this region has at least one putative transmembrane segment or a signal sequence, *i.e.* they are likely to constitute integral membrane proteins, lipoproteins located at the outerside of the membrane, or secreted proteins. These data are not reflective of the findings by Wallin & von Heijne (1998). These researchers reported that 20-30% of the proteins from any genome are predicted to be membrane-localised, and that this fraction increases with genome size. Another general feature that these investigators claimed for membrane proteins from genomes of unicellular organisms, namely the preference for protein

architecture with 6 or 12 transmembrane segments, is only slightly reflected in the region of the *B. subtilis* chromosome studied here. Although, at forehand, this could be an artefact generated by the arbitrary choice of the cut-off value (which is 1 in this study), this is not the case since Wallin & von Heijne have used the same prediction method.

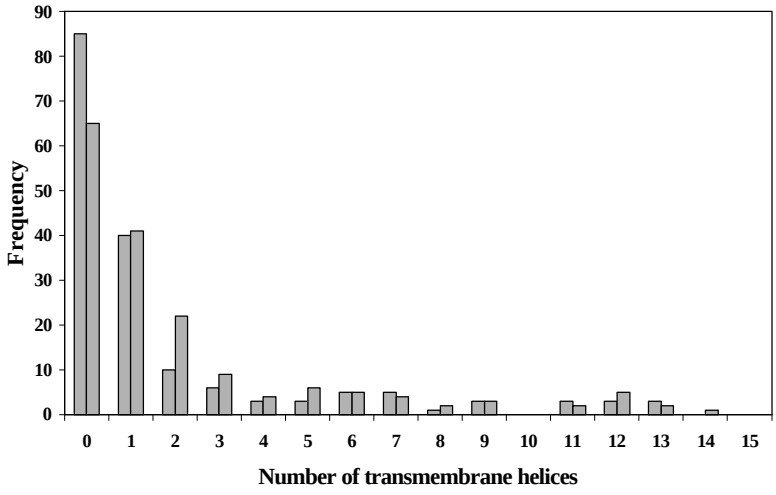


Fig. III.1. Frequency of membrane-spanning domains in proteins from the *prkA-addAB* region (totalling170 proteins). The X-axis represents the number of transmembrane domains, and the Y-axis represents the number of proteins from the region which have this number of domains, as predicted by TopPred2. Left bars represent frequencies with the qualification "certain" according to TopPred2 (score ≥ 1); right bars represent summed-up frequencies with "certain" and "putative" qualifications (score ≥ 0.6).

We also searched the deduced protein sequences of genes from this region for the presence of putative signal sequences characteristic for secreted proteins, and for lipomodification signals. The results of these analyses are listed in tables III.1 and III.2, respectively.

Table III.1. Putative secreted proteins from the *prkA-addAB* region

Name	**Signal sequence**[Ψ] **-1 +1**
AprE	MRSKKLWISLLFALTLIFTMAFSNMS**VQ**AAGKSSTEKKYIVGFKQTMSAM
PhoAIV	MKKMSLFQNMKSKLLPIAAVSVLTAGIF**AG**AELQQTEKASAKKQDKAEIR
YhaK	MRTWKRIPKTTMLISLVSPFLLITPVLFYA**AL**AFPNHAHYFCMISGIHAG
YhaL	MLFFPWWVYLCIVGIIFSAYKL**VA**AAKEEEKVDQAFIEKEGQIYMERMEK
YhdC	MKSLPYTIALLFCGLIIV**SM**AAKGHSTDTDESVQKWEQLAWSKIQDEYKG
YhfM	MKKIVAAIVVIGLVFIAFFYLYSRSGDVYQS**VD**ADLITLSSSGQEDIEIE
YhjA	MKKAAAVLLSLGLVFGFSYGAGHV**AE**AKTKVKVYKNCKELNKVYKGGVAR

[Ψ] Indicated in bold is the cleavage-site for signal peptidase after translocation across the cytoplasmatic membrane. Cleavage occurs at position +1.

Table III.2. Putative lipoproteins in the *prkA-addAB* region

Name	**Signal peptide**	**lipobox**[Ψ] **-3 +1**
PrsA	MKKIAIAAITATSILA**L**SA**C**S
YhaR	MKKVTIAAIHGAAAGLGLS**L**AL**C**A
YhcN	MFGKKQVLASVLLIPLLMTG**C**G

[Ψ] The cysteine at position +1 indicates the site of lipomodification. The preceding amino acids are removed after translocation accross the cytoplasmatic membrane.

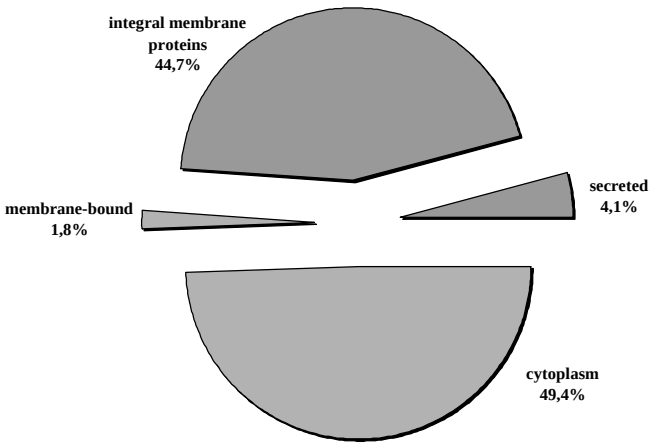


Fig. III.2. Predicted compartmentalisation of proteins encoded by genes from the *prkA-addAB*.

From the information shown in Fig.III.1 and in the Tables III.1 and III.2, the putative localisation of protein from this region can be summarised as shown in Fig.III.2. About 49% of the proteins are cytoplasmatic, 45% inserted into the membrane, 2% membrane-bound through lipomodification, and 4% will be secreted into the surrounding medium.

Paralogs

The availability of the entire *B. subtilis* genome sequence (Kunst *et al*, 1997) enabled us to search on the *B. subtilis* chromosome for paralogs to genes from the *prkA-addAB* region. For this purpose, a paralog was defined as a protein, encoded by the *B. subtilis* chromosome, showing a minimum of 25 % identity over at least three-quarters of the amino acid sequence. The number of paralogous sequences to the ORF products in the region analysed here are listed in Table II.3 (previous chapter), and summarised in Fig. III.3, which depicts the

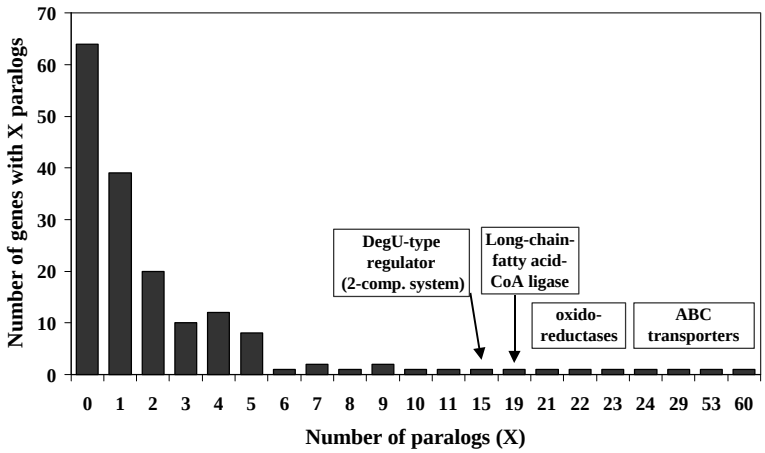


Fig. III.3. Frequency distribution of paralogs of ORFs in the *prkA* to *addAB* region. On the X-axis, the number of paralogs for a given protein sequence is indicated, and on the Y-axis the number of proteins encoded within the *prkA* to *addAB* region for which this number of paralogs is found.

frequency distribution of paralogous sequences to genes from the region studied here. A considerable number of genes in this region have one or more paralogs: only 38 % of the deduced proteins are unique, about 23 % have one paralog, 12 % have two paralogs, etc. Some protein families have very many representatives. For instance, more than 60 members of the ABC transporter family are present on the *B. subtilis* chromosome, with 6

representatives in the region analysed here: *yhaD*, *yhaQ*, *yhcG*, *yhcH*, *yheI*, and *yheH* (see also Fig. III.5). The observed frequency distribution of paralogs from this region is globally similar to that observed for the entire genome (Kunst *et al.*, 1997).

To assess whether some insight in the evolution of the *B. subtilis* chromosome can be deduced from these data, we investigated the chromosomal position of paralogous genes to genes from the *prkA-addAB* region. The results, presented in Fig. III.4, are not indicative of a recent duplication of a large proportion of the chromosome as has been observed in yeast (Goffeau *et al.*, 1996), since no clustering of paralogs could be observed over large regions. However, we did find a number of "paralog hotspots" that may be the result of duplication events on a smaller scale.

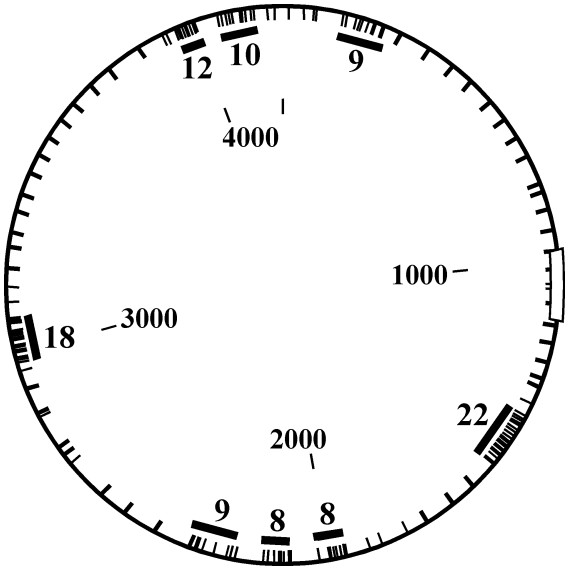


Fig. III.4. Positions of paralogs of ORFs from the *prkA* to *addAB* region. The circle represents the *B. subtilis* chromosome, and co-ordinates are indicated in kilobases relative to the origin of replication. The open box on the circle represents the *prkA-addAB* region. Positions of paralogs are indicated with a dash on the inside of the circle (thin: one paralog; thick: two paralogs). Where paralogs are clustered, this is indicated with a filled bar and the number of paralogs at that position.

The ABC-type transporter proteins, six representatives of which were identified in this region, are exemplary of some of the problems and challenges encountered when performing paralog analyses. This is illustrated in Fig. III.5. According to the definition of paralogs given above, YheH is paralogous to YhaD, but YhaD is not paralogous to YheH. Therefore, it would be desirable to refine the definition of paralogs in order to encompass all significant homologies between proteins. ABC transporters consist of three molecular components or domains, the ATP-binding protein/domain, the membrane protein/domain, and the substrate-binding protein/domain (Tomii & Kanehisa, 1998). By operon- and homology analysis and search for membrane-spanning domains, the functional components of four ABC-like transporters from our region could be identified, and a possible function assigned to them (see also the overview of the region presented in Fig. II.1 and the homology analyses in Table II.3).

The first ABC-like transporter, consisting of YhcG and YhcH, is located in the *yhcE-yhcI* operon with the following components. YhcF, which is a member of the GntR regulator family, is probably the regulator of the operon. YhcG and YhcH both have an ATP-binding domain, but YhcH has an additional domain in the carboxy-terminus. YhcH is homologous to

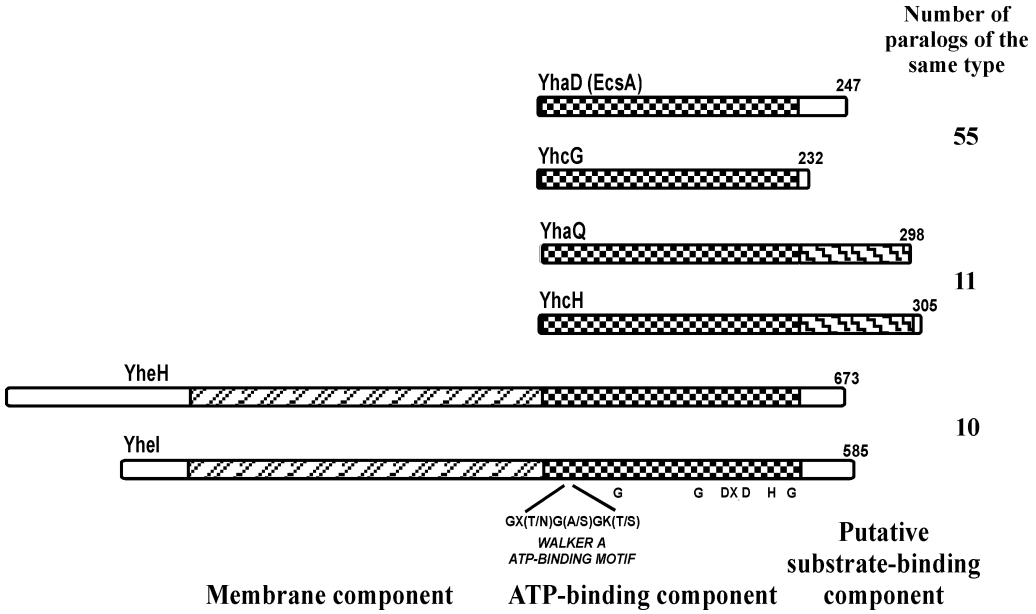


Fig. III.5. Comparison of six proteins that are part of ABC-like transporters in the *prkA-addAB* region of the *B. subtilis* chromosome. The amino-terminal domains present in YheH and YheI (24% identity), contain six membrane-spanning segments ( ). : ATP-binding domains, present in all six proteins. Putative carboxy-terminal substrate-binding domains ( ) were found in YhaQ and YhcH (27% identity). Below the ATP-binding component, the conserved residues including the Walker A ATP binding motif, are indicated. White parts are regions of the proteins without homologies. At the right-hand side of the figure, the number of paralogs of the same type on the *B. subtilis* chromosome is indicated. See text for further details.

the known ABC transporter component BcrA from *Bacillus licheniformis*. YhcI has six membrane-spanning domains (MSDs), and is homologous to BcrB from *B. licheniformis*, which is a component of the same ABC transporter as BcrA. YhcI therefore probably constitutes the membrane component. The C-terminal domain of YhcH may well constitute the substrate-binding domain. The function of YhcE in this operon is unclear although it has six putative MSDs.

The second ABC transporter is encoded by the *yheJ-yheH* operon. YheI and YheH together probably constitute the ATP-binding and membrane components; both contain an ATP-binding domain in the carboxy-terminus and six MSDs in the amino-terminus. Gene *yheI* is preceded by *yheJ*, which shows homology to a domain of phospholipid methyltransferase from *Schizosaccharomyces pombe* (PID:g2209105). This gene may encode the substrate-binding component.

The third ABC transporter is probably encoded by *yhaQ* and *yhaP*, where YhaQ contains the ATP-binding component and, as YhcH, probably the substrate-binding component as well. YhaP has seven MSDs and, therefore, is the most likely candidate for the membrane component. This idea is strengthened by the fact that YhaQ and YhaP are homologous to the *B. subtilis* (and *B. firmus*) Na^+-ABC transporter proteins NatA and NatB, respectively.

The last example is the ABC transporter constituted by YhaD (EcsA), YhaC (EcsB) and YhaB (EcsC). It is involved in exoprotein production, sporulation and competence (Leskela *et*

al., 1996). In this case, *ecsA* encodes the ATP-binding component. EcsB probably constitutes the membrane component, since it contains seven MSDs in a pattern found in other membrane components of ABC transporters. YhaB could constitute the substrate-binding component. However, YhaB also contains two putative MSDs. This protein may therefore also constitute the membrane component since, although some structural similarities can be found, hydrophobic components of ABC transporters do not show extensive amino acid sequence homology.

Evidence for non-functional (remnants of) genes

In the region studied, several ORFs were found of which the deduced proteins are almost certainly not functional or not expressed. This is likely caused by rearrangements and/or deletions within the coding sequence, or the absence of proper transcriptional or translational signals for expression. Based on repeated sequence analysis of these regions, we feel confident that these findings are not the result of sequencing errors.

The first example is the *yzdE-yzdF* pair of partially overlapping ORFs, which code for the N-terminal (*yzdE*) and C-terminal (*yzdF*) fragments of a protein that is present in its entirety in *E. coli* (EmrA; Lomovskaya & Lewis, 1992) and *H. influenzae* (EmrA; Fleischmann *et al.*, 1995). When compared to the *E. coli* and *H. influenzae* genes, the middle 350 bp are absent in *B. subtilis*, which also results in a frameshift (Fig. III.6.). Moreover, *yzdE* is preceded by proper translational start signals (a S-D sequence followed by an ATG start codon), but such signals are absent upstream of *yzdF*.

The second example is *yhaV*. Its deduced ORF product displays significant homology to several HemN proteins, or anaerobic coproporphyrinogen III oxidases involved in heme synthesis under anaerobic conditions, from *H. influenzae* (383 a.a.), *E. coli* (457 a.a.), *Salmonella typhimurium* (457 a.a.), and *Rhodobacter sphaeroides* (305 a.a.). However, no possible translational start site could be found for *yhaV*, and the homology is mainly restricted to the N-terminal two-thirds of the protein.

Another interesting ORF representing a dysfunctional gene, is located upstream of, and partially overlapping with *yhaE*. ORF *yhaE* encodes a possible *B. subtilis* representative of the ubiquitous Hit-like protein (Séraphin, 1992). The first member of this family of proteins was isolated from bovine tissue and identified as being a protein kinase C inhibitor (Pearson *et al.*, 1990). The ORF in front of *yhaE* (results not shown) is 120 codons long, and its deduced a.a. sequence displays blocks of similarity with the catalytic subunit of human DNA-dependent protein kinase (databank ref: PIR|A57099). However, the latter protein is 4096 a.a. long. This finding may be due to “background noise” in the homology search, but the coincidence of finding on ORF with blocks of similarity to a protein kinase, together with a gene encoding a putative protein kinase C inhibitor, is striking.

A similar situation was found downstream of *yhaG*. The deduced YhaG product displays similarity to ProW from *E. coli*, which is involved in a multicomponent binding-protein-dependent transport system for glycine betaine/L-proline (Gowrishankar, 1989). Downstream of *yhaG*, a small ORF was found that shows some similarity to glycine receptor beta subunits from mouse (database reference gp|MMGRBMRA_1), rat (sp|GRB_RAT) and

```
EmrA  H.influenzae  MTQIATENPSTKSVSNKTDRKKGLSIFILLLLIIGIACALYWFFFLKDFEETEDAYVGGN  60
EmrA  E.coli        MSANAETQTPQQPVKKSGKRKRLLLLLTLLFIIIAVAIGIYWFLVLRHFEETDDAYVAGN  60
YzdE  B.subtilis                  MNRGRLILTNIIGLIVVLAIIAGGAYYYYQSTNYVKTDEAKVAGD  45
                                        *     .  *. ..  .*   *..     .  *..* * *.

EmrA  H.influenzae  QVMVSSQVAGNVAKINADNMDKVHAGDILVELDDTNAKLSFEQAKSNLANAVRQVEQLGF  120
EmrA  E.coli        QMQIMSQVSGSVTKVWADNTDFVKEGDVLVTLDPTDARQAFEKAKTALASSVRQTHQLMI  120
YzdE  B.subtilis    MAAITAPAAGKVSDWDLDEGKTVKKGDTVAKIKGEQTVDVKSIMDGTIVKNEVKTDKPYK  105
                       . .  .* *..   *.   *. ** .  ..   .       .  .

EmrA  H.influenzae  TVQQLQSAVHANEISLAQAQGNLARRVQLEKMGAIDKESFQHAKEAVELAKANLNASKNQ  180
EmrA  E.coli        NSKQLQANIEVQKIALAQAQSDYNRRVPLGNANLIGREELQHARDAVTSAQAQLDVAIQQ  180
YzdE  B.subtilis    LVQQLHKRLTWTTYTSQQILKKQILRILK                               134
                      .**.  .     .  *       *.

EmrA  H.influenzae  LAANQALLRNVPLREQPQIQNAINSLKQAWLNLQRTKIRSPIDGYVARRNVQVGQAVSVG  240
EmrA  E.coli        YNANQAMILGTKLEDQPAVQQAATEVRNAWLALERTRIISPMTGYVSRRAVQPGAQISPT  240
YzdF  B.subtilis                                        RKIHYGWHNCEKRSENGQTVQAG  23
                                                                .  . *  .

EmrA  H.influenzae  GALMAVVSNEQMWLEANFKETQLTNMRIGQPVKIHFDLYGKNKEFDGVINGIEMGTGNAF  300
EmrA  E.coli        TPLMAVVPATNMWVDANFKETQIANMRIGQPVTITTDIYGDDVKYTGKVVGLDMGTGSAF  300
YzdF  B.subtilis    TTIAQTIDMDNLYITANIKETDIADIEVGNSVDVVVDGDP-DTTFDGTVEEIGYATNSTF  82
                      .   .   .... ** ***.......*. * .  * .  .    * .  .   *  .*

EmrA  H.influenzae  SLLPSQNATGNWIKVVQRVPVRIKLDPQQFTETPLRIGLSATAKVRISDSSGAMLREKTE  360
EmrA  E.coli        SLLPAQNATGNWIKVVQRLPVRIELDQKQLEQYPLRIGLSTLVSVNTTNRDGQVLANKVR  360
YzdF  B.subtilis    DMLPSTNSSGNYTKVTQKVPVKISIKNPSDKVLPGMNASVKISE                  126
                     .**. *..**. ** *..**.* ..       *

EmrA  H.influenzae  PKTLFSTDTLKYDESAVENLIESIIQQNSHD                               391
EmrA  E.coli        STPVAVSTAREISLAPVNKLIDDIVKANAG                               390
```

Fig III.6. Homology comparison of EmrA amino acid sequences (Multidrug Resistance Protein A) from *E. coli*, *H. influenzae*, and the deduced protein products of *yzdA* and *yzdB* from *B. subtilis*. Amino acid residues that are conserved between *E. coli* and *H. influenzae* are indicated in bold; amino acid residues that are identical in all three organisms are indicated with an asterisk below the three sequences; amino acid residues that are conserved are indicated with a dot.

human (sp:GRB_HUMAN), and an unknown ORF product from *Arabidopsis thaliana*. The deduced ORF product is only 63 a.a. long, while the glycine receptor beta subunits are 484, 496 and 497 a.a. long, respectively, and the similarity is restricted to three small blocks of amino acids. However, a proper S-D sequence with accompanying start codon is present in front of this ORF (AAAGGAGGgagaaggTTG). Functional analysis may reveal the biological relevance of the above mentioned features.

References

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**, 546-567.

Gowrishankar, J. (1989). Nucleotide Sequence of the Osmoregulatory *proU* Operon of *Escherichia coli*. *J.Bacteriol.* **171**, 1923-1931.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Düsterhöft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, H., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., & Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Leskela, S., Kontinen, V. P., & Sarvas, M. (1996). Molecular analysis of an operon in *Bacillus subtilis* encoding a novel ABC transporter with a role in exoprotein production, sporulation and competence. *Microbiology* **142**, 71-77.

Lomovskaya, O. & Lewis, K. (1992). Emr, an *Escherichia coli* locus for multidrug resistance. *Proc.Natl.Acad.Sci.USA* **89**, 8938-8942.

Nielsen, H., Von Heijne, G., & Brunak, S. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**, 1-6.

Pearson, J. D., DeWald, D. B., Mathews, W. R., Mozier, N. M., Zürcher-Neely, H. A., Heinrikson, R. L., Morris, M. A., McCubbin, W. D., McDonald, J. R., Fraser, E. D., Vogel, H. J., Kay, C. M., & Walsh, M. P. (1990). Amino acid sequence and characterization of a protein inhibitor of protein kinase C. *J.Biol.Chem.* **265**, 4583-4591.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc.Natl.Acad.Sci.USA* **85**, 2444-2448.

Séraphin, B. (1992). The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals. *DNA Seq.* **3**, 177-179.

Tomii, K. & Kanehisa, M. (1998). A comparative analysis of ABC transporters in complete microbial genomes. *Genome Research* **10**, 1048-1059.

Von Heijne, G. (1992). Membrane protein structure prediction, hydrophobicity analysis and the positive-inside rule. *J.Mol.Biol.* **225**, 487-494.

Wallin, E. & Von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 1029-1038.

CHAPTER IV

The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*

II.1. Summary

Bacillus subtilis is the best-characterized member of Gram-positive bacteria. Its genome of 4,214,810 base pairs comprises 4,100 protein-coding genes. Of these protein-coding genes, 53% are represented once, while a quarter of the genome corresponds to several gene families that have been greatly expanded by gene duplication, the largest family containing 77 putative ATP-binding transport proteins. In addition, a large proportion of the genetic capacity is devoted to the utilization of a variety of carbon sources, including many plant-derived molecules. The identification of five signal peptidase genes, as well as several genes for components of the secretion apparatus, is important given the capacity of *Bacillus* strains to secrete large amounts of industrially important enzymes. Many of the genes are involved in the synthesis of secondary metabolites, including antibiotics, that are more typically associated with *Streptomyces* species. The genome contains at least ten prophages or remnants of prophages, indicating that bacteriophage infection has played an important evolutionary role in horizontal gene transfer, in particular in the propagation of bacterial pathogenesis.

II.2. Introduction

Techniques for large-scale DNA sequencing have brought about a revolution in our perception of genomes. Together with our understanding of intermediary metabolism, it is now realistic to envisage a time when it should be possible to provide an extensive chemical definition of many living organisms. During the last couple of years, the genome sequences of *Haemophilus influenzae*, *Mycoplasma genitalium*, *Synechocystis* PCC6803, *Methanococcus jannaschii*, *M. pneumoniae*, *Escherichia coli*, *Helicobacter pylori*, *Archaeoglobus fulgidus* and the yeast *Saccharomyces cerevisiae* have been published in their entirety[1-8], and at least 40 prokaryotic genomes are currently being sequenced. Regularly updated lists of genome sequencing projects are available at: http://www.mcs.anl.gov/home/gaasterl/genomes.html (Argonne National Laboratory, IL, USA) and http://www.tigr.org (TIGR, Rockville, MD, USA).

The list of sequenced microorganisms does not currently include a paradigm for Gram-positive bacteria, which are known to be important for the environment, medicine and

industry. *Bacillus subtilis* has been chosen to fill this gap[9,10] as its biochemistry, physiology and genetics have been studied intensely for more than 40 years. *B. subtilis* is an aerobic, endospore-forming, rod-shaped bacterium commonly found in soil, water sources and in association with plants. *B. subtilis* and its close relatives are an important source of industrial enzymes (such as amylases and proteases), and much of the commercial interest in these bacteria arises from their capacity to secrete these enzymes at gram per litre concentrations. It has therefore been used for the study of protein secretion and for development as a host for the production of heterologous proteins[11]. *B. subtilis* (*natto*) is also used in the production of Natto, a traditional Japanese dish of fermented soya beans.

Under conditions of nutritional starvation, *B. subtilis* stops growing and initiates responses to restore growth by increasing metabolic diversity. These responses include the induction of motility and chemotaxis, and the production of macromolecular hydrolases (proteases and carbohydrases) and antibiotics. If these responses fail to re-establish growth, the cells are induced to form chemically, irradiation and desiccation resistant endospores. Sporulation involves a perturbation of the normal cell cycle and the differentiation of a binucleate cell into two cell types. The division of the cell into a smaller forespore and a larger mother cell, each with an entire copy of the chromosome, is the first morphological indication of sporulation. The former is engulfed by the latter and differential expression of their respective genomes, coupled to a complex network of interconnected regulatory pathways and developmental checkpoints, culminates in the programmed death and lysis of the mother cell and release of the mature spore[12]. In an alternative developmental process, *B. subtilis* is also able to differentiate into a physiological state, the competent state, that allows it to undergo genetic transformation[13].

IV.3. General features of the DNA sequence

Analysis at the replicon level.

The *B. subtilis* chromosome has 4,214,810 base pairs (bp), with the origin of replication coinciding with the base numbering start point[14], and the terminus at about 2,017 kilobases (kb)[15]. The average G+C ratio is 43.5%, but it varies considerably throughout the chromosome. This average is also different if one considers the nucleotide content of coding sequences, for which G and A (24% and 30%) are relatively more abundant than their counterparts C and T (20% and 26%). A significant inversion of the relative G-C/G+C ratio is visible at the origin of replication, indicating asymmetry of the nucleotide composition between the replication leading strand and the lagging strand[16]. Several A+T-rich islands are likely to reveal the signature of bacteriophage lysogens or other inserted elements (Fig. IV.1, see also below).

We have analysed the abundance of oligonucleotides (words) in the genome in various ways: absolute number of words in the genomic text, or comparison with the expected count derived from several models of the chromosome (for example, Markov models, or simulated sequences in which previously known features of the genome were conserved[17]). Comparing the experimental data with various models allowed us to define under- and overrepresentation

of words in the experimental data set by reference to the model chosen. In general, the dinucleotide bias follows closely what has been described for other prokaryotes[18,19], in that the

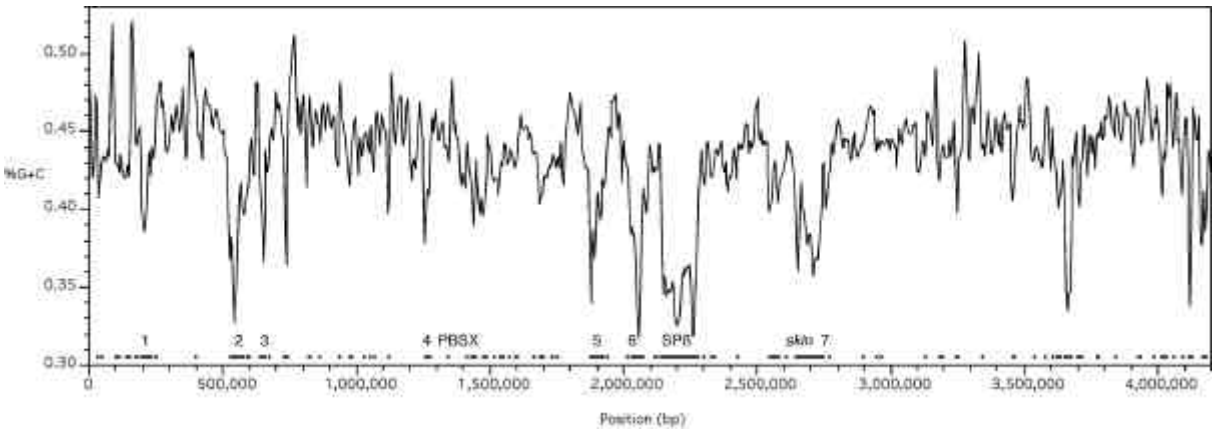


Figure IV.1. Distribution of A+T-rich islands along the chromosome of *B. subtilis*, in sliding windows of 10,000 nucleotides, with a step of 5,000 nucleotides. Location of genes from class 3 (see text and Fig. 4) is indicated by dots at the bottom of the graph. Known prophages (PBSX, SPβ, *skin*) are indicated by their names, and prophage-like elements are numbered from 1 to 7.

dinucleotides most overrepresented are AA, TT and GC, whereas those less represented are TA, AC and GT. Plots of the frequencies of AG, GA, CT and TC in sliding windows along the chromosome show dramatic decreases or increases around the origin and terminus of replication (data not shown). Trinucleotide frequency, directly related to the coding frame, will be discussed below. The distribution of words of four, five and six nucleotides shows significant correlations between the usage of some words and replication (several such oligonucleotides are very significantly overrepresented in one of the strands and underrepresented in the other one).

Setting a statistical cut-off for the significance of duplications at 10^{-3}, we expected duplication by chance of words longer than 24 nucleotides to be rare[20]. In fact, the genome of *B. subtilis* contains a plethora of such duplications, some of them appearing more than twice. Among the duplications, we identified, as expected, the ribosomal RNA genes and their flanking regions, but also regions known to correspond to genes comprising long sequence repeats (such as *pks* and *srf*). We also found several regions that were not expected: a 182 bp repetition within the *yyaL* and *yyaO* genes; a 410 bp repetition between the *yxaK* and *yxaL* genes; an internal duplication of 174 bp inside *ydcI*; and significant duplications in the regions involved in the transcriptional control of several genes (such as 118 bp repeated three times between *yxbB* and *yxbC*). Finally, we found several repetitions at the borders of regions that might be involved in bacteriophage integration.

The most prominent duplication was a 190 bp element that was repeated 10 times in the chromosome. Multiple alignment of the ten repeats showed that they could be classified into two subfamilies with six and three copies each, plus a copy of what appears to be a chimaera. Similar sequences have also been described in the closely related species, *Bacillus licheniformis*[21,22]. A striking feature of these repeats is that they are only found in half of the

chromosome, at either side of the origin of replication, with five repeats on each side. Furthermore, with the exception of the most distal repeat at position 737,062, they lie in the same orientation with respect to the movement of the replication fork (Figs. IV.2♦ and IV.3). Putative secondary structures conserved by compensatory mutations, as well as an insert in three of the copies, suggest that this element could indicate a structural RNA molecule.

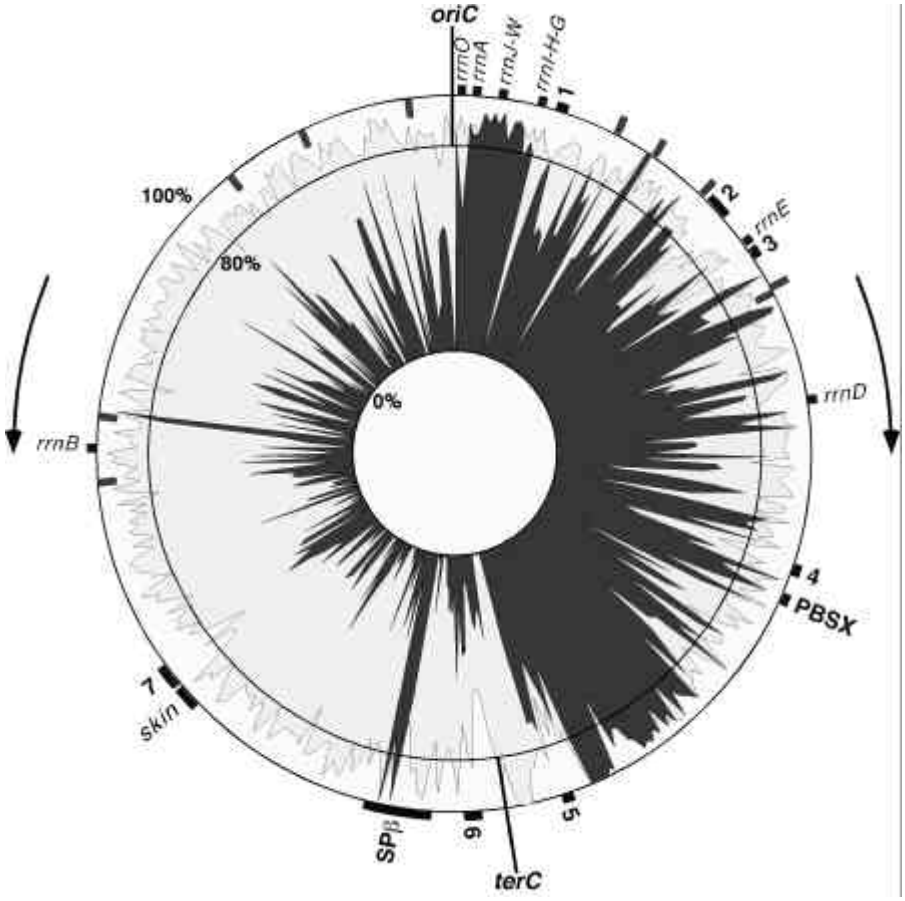


Fig. IV.3. Density of coding nucleotides along the *B. subtilis* chromosome. Light-grey stands for the density of coding nucleotides in both strands of the sequence, whereas dark-grey indicates the density of coding nucleotides in the clockwise strand (*i.e.* nucleotides involved in genes transcribed in the clockwise orientation). The movement of the replication forks is represented by arrows. Ribosomal RNA operons are indicated (*rrn*). Known prophages and prophage-like elements are represented as dark lines at the outside of the circle. The 190 bp element repeated ten times is represented by sticks.

Analysis at the transcription and translation level

Over 4,000 putative protein coding sequences (CDSs) have been identified, with an average size of 890 bp, covering 87% of the genome sequence (Fig. IV. 2). We found that

♦ Fig IV.2 can be accessed at: http://www.pasteur.fr/BIO/SubtiList.html

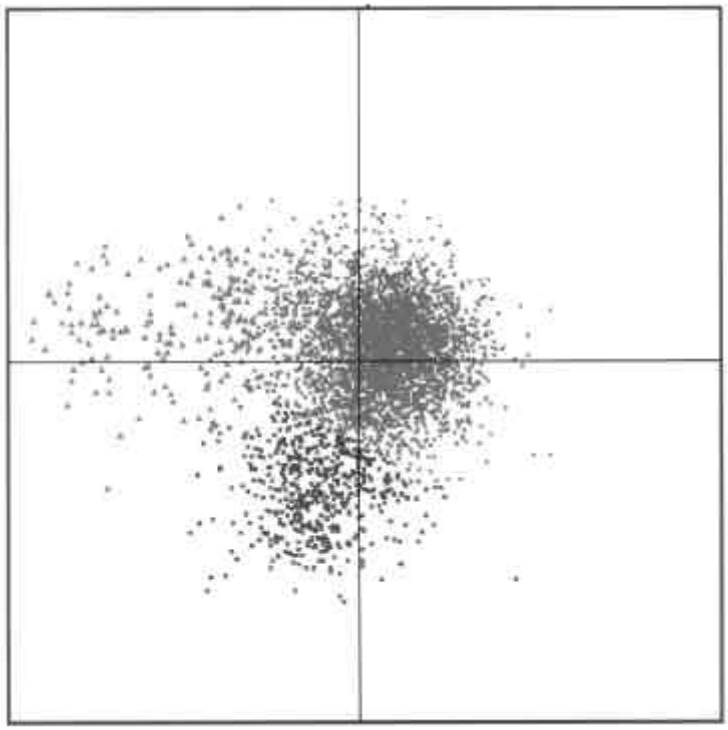


Fig. IV.4. Factorial Correspondence Analysis of codon usage in the *B. subtilis* CDSs. Genes from class 1 are represented by squares, genes from class 2 are represented by triangles, and genes from class 3 are represented by crosses. Class 2 contains genes coding for the translation and transcription machineries, and genes of the core intermediary metabolism. Class 3 genes correspond to codons strongly enriched in A or T in the wobble position. They generally belong to prophage-like inserts in the genome.

related *Bacillus* species, especially as strain 168 is derived from the Marburg strain that was subjected to X-ray irradiation[23].

A few regions do not have any identifiable feature indicating that they are transcribed: they could be "grey holes" of the type described in *E. coli*[24]. Preliminary studies involving all regions of more than 400 bp without annotated CDSs indicated that, of ~300 such regions, only 15% were likely to be really devoid of protein-coding sequences. One of the longest such regions, located between *yfjO* and *yfjN*, is 1,628 bp long. Grey holes seem generally to be clustered near the terminus of replication. However, a grey-hole cluster located at ~600 kb might be related to the temporary chromosome partition observed during the first stages of sporulation, when a segment of about 1/3 of the chromosome enters the prespore, and remains the sole part of the chromosome in the prespore for a significant transition period[25].

The codon usage of *B. subtilis* CDSs was analysed using factorial correspondence analysis[17]. We found that the CDSs of *B. subtilis* could be separated into three well-defined classes (Fig. IV.4). Class 1 forms the majority of the *B. subtilis* genes (3,375 CDSs), including most of the genes involved in sporulation. Class 2 (188 CDSs) includes genes that are highly expressed under exponential growth conditions, such as genes encoding the transcription and translation machineries, core intermediary metabolism, stress proteins, and one third of genes of unknown function. Class 3 (537 CDSs) contains a very high proportion of genes of unknown function (84%), and the members of this class have codons enriched in A+T residues. These genes are usually clustered into groups of 15 to 160 genes (for example, bacteriophage SPß) and correspond to the A+T-rich islands described above (Fig. IV.1). When they are of known function, or when their products display similarity to proteins of known function, they usually correspond to functions found in, or associated with, bacteriophages or transposons, as well as functions related to the cell envelope. This includes the region *ydc/ydd/yde* (40 genes that are missing in some *B. subtilis* strains[26]), where gene products showing similarities to

bacteriophage and transposon proteins are intertwined. Many of these genes are associated with virulence genes identified in pathogenic Gram-positive bacteria, suggesting that such virulence factors are transmitted horizontally among bacteria at a much higher frequency than previously thought. If we include these A+T-rich regions as possible cryptic phages, together with known bacteriophages or bacteriophage-like elements (SPß, PBSX and the *skin* element), we find that the genome of *B. subtilis* 168 contains at least 10 such elements (Figs. IV.2 and IV.3). Annotation of the corresponding regions often reveals the presence of genes that are similar to bacteriophage lytic enzymes, perhaps accounting for the observation that *B. subtilis* cultures are extremely prone to lysis.

The ribosomal RNA genes have been previously identified and shown to be organized into ten rRNA operons, mainly clustered around the origin of replication of the chromosome (Figs. IV.2 and IV.3). In addition to the 84 previously identified tRNA genes, by using the Palingol[27] and tRNAscan[28] programs, we propose four putative new tRNA loci (at 1,262 kb, 1,945 kb, 2,003 kb, 2,899 kb), specific for lysine, proline and arginine (UUU, GGG, CCU and UCU anticodons, respectively). The 10S RNA involved in degradation of proteins made from truncated mRNA has been identified (*ssrA*), as well as the RNA component of RNase P (*rnpB*) and the 4.5S RNA involved in the secretion apparatus (*scr*).

There is a strong transcription orientation bias with respect to the movement of the replication fork: 75% of the predicted genes are transcribed in the direction of replication. Plotting the density of coding nucleotides in each strand along the chromosome readily identifies the replication origin and terminus (Fig. IV.3). To identify putative operons, we followed ref. 29 for describing Rho independent transcription termination sites. This yielded *ca* 1,630 putative terminators (340 of which were bidirectional). We retained only those that were located less than 100 bp downstream of a gene, or that were considered by the program to be "very strong" (in order to account for possible erroneous CDSs). This yielded a total of ~1,250 terminators, with a mean operon size of 3 genes. A similar approach to the identification of promoters is problematical, especially because at least 14 sigma factors, recognising different promoter sequences, have been identified in *B. subtilis*. Nevertheless, the consensus of the main vegetative sigma factor (σ^A) appears to be identical to its counterpart in *E. coli* (σ^{70}): 5'-TTGACA-n_{17}-TATAAT. Relaxing the constraints of the similarity to sigma-specific consensus sequences led to an extremely high number of false-positive results, suggesting that the consensus-oriented approach to the identification of promoters should be replaced by another approach[17].

IV.4. Classification of gene products

Genes were classified according to ref. 14, based on the representation of cells as Turing machines in which one distinguishes between the machine and the program (Table IV.1♦). Using the BLAST2P software running against a composite protein databank compound of

♦ Table IV.1 can be accessed at: http://www.pasteur.fr/BIO/SubtiList.html

SWISS-PROT (release 34), TREMBL (release 3, update 1), and *B. subtilis* proteins, we assigned at least one significant counterpart with a known function to 58% of the *B. subtilis* proteins. Thus for up to 42% of the gene products, the function cannot be predicted by similarity to proteins of known function: 4% of the proteins are similar only to other unknown proteins of *B. subtilis*; 12% are similar to unknown proteins from some other organism; and 26% of the proteins are not significantly similar to any other proteins in databanks. This preliminary analysis should be interpreted with caution, because only ~1,200 gene functions (30%) have been experimentally identified in *B. subtilis*. We used the '*y*' prefix in gene names to emphasize that the function has not been ascertained (2,853 '*y*' genes, representing 70%).

Regulatory systems

Transcription regulatory proteins. Helix-Turn-Helix proteins form a large family of regulatory proteins found in both prokaryotes and eukaryotes. There are several classes, including repressors, activators and sigma factors. Using BLAST searches, we constructed consensus matrices for helix-turn-helix proteins to analyse the *B. subtilis* protein library. We identified 18 sigma or sigma-like factors, of which nine (including a new one) are of the SigA type. We also putatively identified 20 regulators (among which 18 were products of '*y*' genes) of the GntR family, 19 regulators (15 '*y*' genes) of the LysR family, and 12 regulators (5 '*y*' genes) of the LacI family. Other transcription regulatory proteins were of the AraC family (11 members, 10 '*y*'), the Lrp family (7 members, 3 '*y*'), the DeoR family (6 members, 3 '*y*'), or additional families (such as the MarR, ArsR, or TetR families). A puzzling observation is that several regulatory proteins display significant similarity to aminotransferases (seven such enzymes have been identified as showing similarity to repressors).

Two-component signal-transduction pathways

Two-component regulatory systems, consisting of a sensor protein kinase and a response regulator, are widespread among prokaryotes. We have identified 34 genes encoding response regulators in *B. subtilis*, most of which having adjacent genes encoding histidine kinases. Response regulators possess a well-conserved N-terminal phospho-acceptor domain[30], whereas their C-terminal DNA binding domains share similarities with previously identified response regulators in *E. coli*, *Rhizobium meliloti*, *Klebsiella pneumoniae* or *Staphylococcus aureus.* Representatives of the four subfamilies recently identified in *E. coli*[31] (OmpR, FixJ, CitB and LytR) have been identified in *B. subtilis*. In a fifth subfamily, CheY, the DNA binding domain is absent. The DNA binding domain of a single *B. subtilis* response regulator, YesN, shares similarity with regulatory proteins of the AraC family.

Quorum sensing

The *B. subtilis* genome contains 11 aspartate phosphatase genes, whose products are involved in dephosphorylation of response regulators, that do not seem to have counterparts in Gram-negative bacteria such as *E. coli*. Downstream from the corresponding genes are some

small genes, called *phr*, encoding regulatory peptides which may serve as quorum sensors[32]. Seven *phr* genes have been identified so far, including three new genes (*phrG*, *phrI* and *phrK*).

Protein secretion

It is known that *B. subtilis* and related *Bacillus* species, in particular *B. licheniformis* and *B. amyloliquefaciens*, have a high capacity to secrete proteins into the culture medium. Several genes encoding proteins of the major secretion pathway have been identified: *secA*, *secD*, *secE*, *secF*, *secY*, *ffh* and *ftsY*. Surprisingly, there is no gene for the SecB chaperone. It is thought that other chaperone(s) and targeting factor(s), such as Ffh and FtsY, may take over the SecB function. Further, although there is only one such gene in *E. coli*, five type I signal peptidase genes (*sipS*, *sipT*, *sipU*, *sipV* and *sipW*, have been found[33]. The *lsp* gene, encoding a type II signal peptidase required for processing of lipo-modified precursors, was also identified. PrsA, located at the outer side of the membrane, is important for the refolding of several mature proteins after their translocation through the membrane.

Other families of proteins

ABC transporters were the most frequent class of proteins found in *B. subtilis*. They must be extremely important in Gram-positive bacteria, because they have an envelope comprising a single membrane. ABC transporters will therefore allow such bacteria to escape the toxic action of many compounds. We propose that 77 such transporters are encoded in the genome. In general they involve the interaction of at least 3 gene products, specified by genes organised into an operon. Other families comprised 47 transport proteins similar to facilitators (and perhaps sometimes part of the ABC transport systems), 18 amino acid permeases (probably antiporters), and at least 16 sugar transporters belonging to the PEP-dependent phosphotransferase system.

General stress proteins are important for the survival of bacteria under a variety of environmental conditions. We identified 43 temperature-shock and general stress proteins displaying strong similarity to *E. coli* counterparts.

Missing genes

Histone-like proteins such as HU and H-NS have been identified in *E. coli*. We found that *B. subtilis* encodes two putative histone-like proteins that show similarity to *E. coli* HU, namely HBsu and YonN, but found no homologue to H-NS. It is known that the *hbs* gene encoding HBsu is essential, but we do not expect the *yonN* gene to be essential because it is present in the SPß prophage. IHF is similar to HU, and it is not known whether HBsu plays a similar role to that of IHF in *E. coli*. Similarly, no protein similar to FIS could be found.

Genes encoding products that interact with methylated DNA, such as *seqA* in *E. coli*, involved in the regulation of replication initiation timing, or *mutH*, the endonuclease recognizing the newly synthesized strand during mismatch repair at hemi-methylated GATC sites, are also missing. This is in line with the absence of known methylation in *B. subtilis*, equivalent to Dam methylation in *E. coli*. Similarly, *E. coli sfiA*, encoding an inhibitor of FtsZ

action in the SOS response, has no counterpart in *B. subtilis*. In contrast, *B. subtilis* replication initiation-specific genes, such as *dnaB* and *dnaD*, are missing in *E. coli*. The exact counterpart of the *E. coli mukB* gene, involved in chromosome partitioning, does not exist in *B. subtilis*, but genes *spo0J* and *smc* (Smc is weakly similar to MukB), which are suggested to be involved in partitioning of *B. subtilis* chromosome, are missing in *E. coli*.

Turnover of mRNA is controlled in *E. coli* by a 'degradosome' comprising RNase E. It has a counterpart in *B. subtilis*, but we failed to find a clear homologue of RNase E in this organism. Whether this is related to the role of ribosomal protein S1 as an RNA helicase involved in mRNA turnover in *E. coli* requires further investigation. In particular, a homologue of *rpsA* (S1 structural gene), *ypfD*, might be involved in a structure homologous to the degradosome[34].

Structurally unrelated genes of similar function

Several genes encode products that have similar functions in *E. coli* and *B. subtilis*, but have no evident common structure. This is the case of the helicase loader genes: *E. coli dnaC* and *B. subtilis dnaI*; the genes coding for the replication termination protein, *E. coli tus* and *B. subtilis rtp*; and the division topology specifier genes: *E. coli minE* and *B. subtilis divIVA*. The situation may even be more complex in multisubunit enzymes: *B. subtilis* synthesizes two DNA polymerase III α chains, one having 3'-5' proofreading exonuclease activity (PolC) and the other one without the exonuclease activity (DnaE); in *E. coli*, only the latter exists. *E. coli* DNA polymerase II is structurally related to DNA polymerase α of eukaryotes, whereas *B. subtilis* YshC is related to DNA polymerase ß.

IV.5. Metabolism of small molecules

The type and range of metabolism used for the interconversion of low-molecular-weight compounds provide important clues to an organism's natural environment(s) and its biological activity. Here we briefly outline the main metabolic pathways of *B. subtilis* before the reconstruction of these pathways *in silico*, the correlation of genes with specific steps in the pathway, and ultimately the prediction of patterns of gene expression.

Intermediary metabolism

It has long been known that *B. subtilis* can use a variety of carbohydrates. As expected, it encodes an Embden-Meyerhof-Parnas glycolytic pathway, coupled to a functional tricarboxylic acid cycle. Further, *B. subtilis* is also able to grow anaerobically in the presence of nitrate as an electron acceptor. This metabolism is, at least in part, regulated by the FNR protein, binding to sites upstream of at least eight genes (four sites experimentally confirmed and four putative sites). A noteworthy feature of *B. subtilis* metabolism is an apparent requirement of branched short-chain carboxylic acids for lipid biosynthesis[35]. Branched-chain 2-keto acid decarboxylase activity exists and may be linked to a variety of genes, suggesting

that *B. subtilis* can synthesize and utilize linear branched short-chain carboxylic acids and alcohols.

Amino-acid and nucleotide metabolism

Pyrimidine metabolism of *B. subtilis* seems to be regulated in a way fundamentally different from that of *E. coli*, as it has two carbamylphosphate synthetases (one specific for arginine synthesis, the other for pyrimidine). Additionally, the aspartate transcarbamylase of *B. subtilis* does not act as an allosteric regulator as it does in *E. coli*. As in other microorganisms, pyrimidine deoxyribonucleotides are synthesized from ribonucleoside diphosphates, not triphosphates. The cytidine diphosphate required for DNA synthesis is derived from either the salvage pathway of mRNA turnover or from the synthesis of phospholipids and components of the cell wall. This means that polynucleotide phosphorylase is of fundamental importance in nucleic acid metabolism, and may account for its important role in competence[36]. Two ribonucleoside reductases, both of Class I, NrdEF type, are encoded by the *B. subtilis* chromosome, in one case from within the SPβ genome. In this latter case, the gene corresponding to the large subunit both contains an intron and codes for an intein (V.L., unpublished data). The gene of the small subunit of this enzyme also contains an intron, encoding an endonuclease, as was found for the homologue in bacteriophage T4.

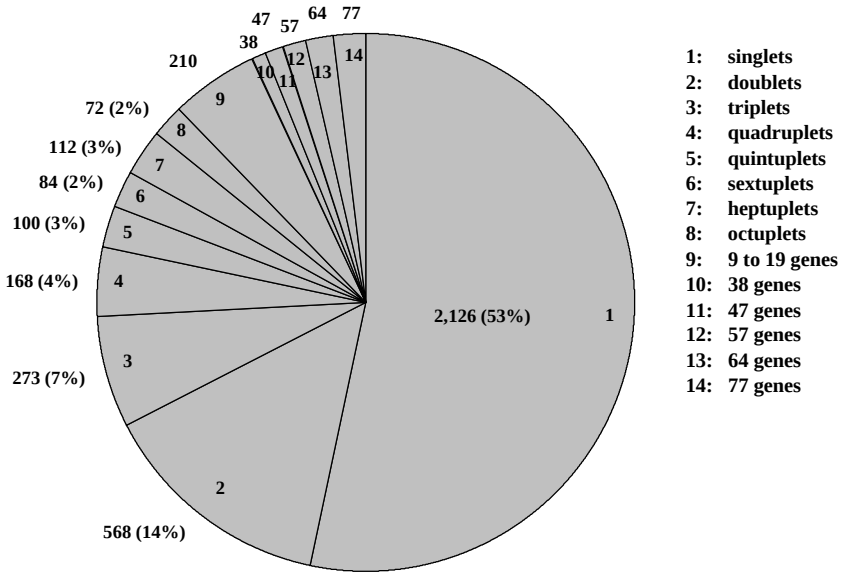
By similarity with genes from other organisms, there appears to be, in addition to genes involved in amino-acid degradation (such as the *roc* operon, which degrades arginine and related amino acids), a large number of genes involved in the degradation of molecules such as opines and related molecules, derived from plants. This is also in line with the fact that *B. subtilis* degrades polygalaturonate, and suggests that, in its biotope, it forms specific relations with plants.

Secondary metabolism

In addition to many genes coding for degradative enzymes, almost 4% of the *B. subtilis* genome codes for large multifunctional enzymes (for example, the *srf*, *pps* and *pks* loci), similar to those involved in the synthesis of antibiotics in other genera of Gram-positive bacteria such as *Streptomyces*. Natural isolates of *B. subtilis* produce compounds with antibiotic activity, such as surfactin, fengycin and difficidin, that can be related to the above-mentioned loci. This bacterium therefore provides a simple and genetically amenable model in which to study the synthesis of antibiotics and its regulation. These pathways are often organized in very long operons (for example, the *pks* region spans 78.5 kb, about 2% of the genome). The corresponding sequences are mostly located near the terminus of replication, together with prophages and prophage-like sequences.

IV.6. Paralogues and orthologues

It is important to relate intermediary metabolism to genome structure, function and evolution. We have therefore compared the *B. subtilis* proteins with themselves, as well as with proteins from known complete genomes, using a consistent statistical method that permits



evaluation of unbiased probabilities of similarities between proteins[37,38]. For Z-scores higher than 13, the number of proteins similar to each given protein does not vary, indicating that this cut-off value identifies sets of proteins that are significantly similar.

Families of paralogues

Many of the paralogues constituted large families of functionally related proteins, involved in the transport of compounds into and out of the cell, or involved in transcription regulation. Another part of the genome consisted of gene doublets (568 genes), triplets (273 genes), quadruplets (168 genes) and quintuplets (100 genes). Finally, about half of the genome is made of genes coding for proteins with no apparent paralogues (Fig. IV.5). No large family comprises only proteins without any similarity to proteins of known function.

The process by which paralogs are generated is not well understood, but we might find clues by studying some of the duplications in the genome. Several approximate DNA repetitions, associated with very high levels of protein identity, were found, mainly within regions putatively or previously identified as prophages. This is in line with previous observations about PBSX and the *skin* element[39,40], and suggests that these prophage-like elements share a common ancestor and have diverged relatively recently. In addition, several protein duplications are in genes that are located very close to each other, such as *yukL* and *dhbF* (the corresponding proteins are 65% identical in an overlap of 580 amino acids), *yugJ* and *yugK* (proteins 73% identical), *yxjG* and *yxjH* (proteins 70% identical), and the entire *opuB* operon which is duplicated 3 kb away (*opuC* operon, yielding ~80% of amino-acid identity in the corresponding proteins).

The study of paralogues showed that, as in other genomes, a few classes of genes have been highly expanded. This argues against the idea of the genome evolving through a series of duplications of ancestral genomes, but rather for the idea of genes as living organisms, subjected to evolutionary constraints, some being submitted to expansion and natural selection, and others to local duplications of DNA regions.

Among paralogue doublets, some were unexpected, such as the three aminoacyl tRNA synthetases doublets (*hisS* (2,817 kb) and *hisZ* (3,588 kb); *thrS* (2,960 kb) and *thrZ* (3,855 kb); *tyrS* (3,036 kb) and *tyrZ* (3,945 kb)) or the two *mutS* paralogues (*mutS* and *yshD*). This latter situation is similar to that found in *Synechocystis*. In the case of *B. subtilis*, the presence of two MutS proteins could indicate that there are two different pathways for long-patch mismatch repair, possibly a consequence of the active genetic transformation mechanism of *B. subtilis*.

Families of orthologues

Because *Mycoplasma* spp. are thought to be derived from Gram-positive bacteria similar

Fig. IV.5 Gene paralog distribution in the genome of *B. subtilis*. Each *B. subtilis* protein has been compared to all other proteins in the genome, using a Smith and Waterman algorithm. The baseline is established by making a similar comparison using 100 independent random shuffles of the protein sequence (Z-score > 13).

to *B. subtilis*, we compared the *B. subtilis* genome to that of *M. genitalium*. Among the 450 genes encoded by *M. genitalium*, the products of 300 are similar to proteins of *B. subtilis*. Among the 146 remaining gene products, a further 3 are similar to proteins of other *Bacillus* species, and 9 to proteins of other Gram-positive bacteria; 25 are similar to proteins of Gram-negative bacteria; and 19 are similar to proteins of other *Mycoplasma* spp. This leaves only 90 genes that would be specific to *M. genitalium* and might be involved in the interaction of this organism with its host.

The *B. subtilis* genome is similar in size to that of *E. coli*. Because these bacteria probably diverged more than one billion years ago, it is of evolutionary value to investigate their relative similarity. About 1,000 *B. subtilis* genes having clear orthologous counterparts in *E. coli* (one-quarter of the genome). These genes did not belong either to the prophage-like regions or to regions coding for secondary metabolism (~15% of the *B. subtilis* genome). This indicates that a large fraction of these genomes shared similar functions. At first sight, however, it seems that little of the operon structure has been conserved. We nevertheless found that ~100 putative operons or part of operons were conserved between *E. coli* and *B. subtilis*. Among these, ~12 exhibited a reshuffled gene order (typically, the arabinose operon is *araABD* in *B. subtilis* and *araBAD* in *E. coli*). In addition to the core of the translation and transcription machinery, we identified other classes of operons that were well conserved between the two organisms, including major integrated functions such as ATP synthesis (*atp* operon) and electron transfer (*cta* and *qox* operons). As well as being well preserved, the murein biosynthetic region was partly duplicated, allowing creation of part of the genes required for the sporulation division machinery[41]. The amino-acid biosynthesis genes

differ more in their organisation: the *E. coli* genes for arginine biosynthesis are spread throughout the chromosome; whereas the arginine biosynthesis genes of *B. subtilis* form an operon. The same is true for purine biosynthetic genes. Genes responsible for the biosynthesis of coenzymes and prosthetic groups in *B. subtilis* are often clustered in operons that differ from those found in *E. coli*. Finally, several operons conserved in *E. coli* and *B. subtilis* correspond to unknown functions, and should therefore be priority targets for functional analysis of these model genomes.

Comparison with *Synechocystis* PCC6803 revealed about 800 orthologs. However, in this case the putative operon structure is extremely poorly conserved, apart from four of the ribosomal protein operons, the *groES-groEL* operon, *yfnHG* (respectively in *Synechocystis rfbFG*), *rpsB-tsf, ylxS-nusA-infB, asd-dap-ymfA, spmAB, efp-accB, grpE-dnaK, yurXW*. The nine-gene *atp* operon of *B. subtilis* is split into two parts in *Synechocystis*: *atpBE* and *atpIHGFDAC*.

IV.7. Conclusion

The biochemistry, physiology and molecular biology of *B. subtilis* have been extensively studied over the past 40 years. In particular, *B. subtilis* has been used to study postexponential phase phenomena such as sporulation and competence for DNA uptake. The genome sequences of *E. coli* and *B. subtilis* provide a means of studying the evolutionary divergence, one billion years ago, of eubacteria into the Gram-positive and Gram-negative groups.

The availability of powerful genetic tools will allow the *B. subtilis* genome sequence data to be exploited fully within the framework of a systematic functional analysis program, undertaken by a consortium of 19 European and 7 Japanese laboratories coordinated by S. D. Ehrlich (INRA, Jouy-en-Josas, France) and by N. Ogasawara and H. Yoshikawa (Nara Institute of Science and Technology, Nara, Japan).

IV.8. Methods

Genome cloning and sequencing

An international consortium was established to sequence the genome of *B. subtilis* strain 16842. At its peak, 25 European, seven Japanese and one Korean laboratory participated in the program, together with two biotechnology companies. Five contiguous DNA regions totalling 0.94 Mb, and two additional regions of 0.28 and 0.14 Mb, were sequenced by the Japanese partners, while the European partners sequenced a total of 2.68 Mb. A few sequences from strain 168 published previously were not resequenced when long overlaps did not indicate differences.

A major technical difficulty was the inability to construct in *E. coli* gene banks representative of the entire *B. subtilis* chromosome using vectors that have proved efficient for other sources of bacterial DNA (such as bacteriophage or cosmid vectors). This was due to the generally very high level of expression of *B. subtilis* genes in *E. coli*, leading to toxic effects. This limitation was overcome by: cloning into a variety of vectors[9,43,44]; using an *E. coli* strain

maintaining low-copy number plasmids[44]; using an integrative plasmid/marker rescue genome walking strategy[44]; and *in vitro* amplification using polymerase chain reaction (PCR) techniques[45,46].

Although cloning vectors were used in the early stages as templates for sequencing reactions, they were largely superseded in the later stages by long-range and inverse PCR techniques. To reduce sequencing errors resulting from PCR amplification artefacts, at least eight amplification reactions were performed independently and subsequently pooled. The various sequencing groups were free to choose their own strategy, except that all DNA sequences had to be determined entirely on both strands.

Sequence annotation and verification

The sequences were annotated by the groups, and sent to a central depository at the Institut Pasteur[14]. The Japanese sequences were also sent there through the Japanese depository at the Nara Institute of Science and Technology. The same procedures were used to identify CDSs and to detect frameshifts. They were embedded within a cooperative computer environment dedicated to automatic sequence annotation and analysis[39]. In a first step, we identified in all six possible frames the open reading frames (ORFs) that were at least 100 codons in length. In a second step, three independent methods were used: the first method used the GeneMark coding-sequence prediction method[47] together with the search for CDSs preceded by typical translation initiation signals (5'-AAGGAGGTG-3'), located 4-13 bases upstream of the putative start codons (ATG, TTG or GTG); the second method used the results of a BLAST2X analysis performed on the entire *B. subtilis* genome against the non-redundant protein databank at the NCBI; and the third method was based on the distribution of non-overlapping trinucleotides or hexanucleotides in the three frames of an ORF[48].

In general, frameshifts and missense mutations generating termination codons or eliminating start codons are relatively easy to detect. We shall devise a procedure for detecting another type of error, GC instead of CG or vice versa, which are much more difficult to identify. It should be noted that putative frameshift errors should not be corrected automatically. The sequences of the flanking regions of a 500-bp fragment centred around a putative error were sent to an independent verification group, which performed PCR amplifications using chromosomal DNA as template, and sequenced the corresponding DNA products.

Organization and accessibility of data

The *B. subtilis* sequence data have been combined with data from other sources (biochemical, physiological and genetic) in a specialized database, SubtiList[49], available as a Macintosh or Windows stand-alone application (4th Dimension runtime) by anonymous ftp at ftp://ftp.pasteur.fr/pub/GenomeDB/SubtiList. SubtiList is also accessible through a World-Wide Web server at http://www.pasteur.fr/Bio/SubtiList.html, where it has been implemented on a UNIX system using the Sybase relational database management system. A completely rewritten version of SubtiList is in preparation to facilitate browsing of the information of the whole chromosome. Flat files of the whole DNA and protein sequences in EMBL and FASTA

format will be made available at the above ftp address. Another *B. subtilis* genome database is also under development at the Human Genome Center of Tokyo University (http://www.genome.ad.jp), and SubtiList will also be available there.

References

1. Kunst, F., Vassarotti, A. & Danchin, A. Organization of the European *Bacillus subtilis* genome sequencing project. *Microbiology* **389**, 84-87 (1995).
2. Ogasawara, N. & Yoshikawa, H. The systematic sequencing of the *Bacillus subtilis* genome in Japan. *Microbiology* **142**, 2993-2994 (1996).
3. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).
4. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403 (1995).
5. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109-136 (1996).
6. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073 (1996).
7. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420-4449 (1996).
8. Goffeau, A. *et al.* The yeast genome directory. *Nature* **387**, 5-105 (1997).
9. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547 (1997).
10. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462 (1997).
11. Harwood, C. R. *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* **10**, 247-256 (1992).
12. Stragier, P. & Losick, R. Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.* **30**, 297-341 (1996).
13. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet.* **12**, 150-155 (1996).
14. Moszer, I., Kunst, F. & Danchin, A. The European *Bacillus subtilis* genome sequencing project: current status and accessibility of the data from a new World Wide Web site. *Microbiology* **142**, 2987-2991 (1996).
15. Franks, A. H., Griffiths, A. A. & Wake, R. G. Identification and characterization of new DNA replication terminators in *Bacillus subtilis*. *Mol. Microbiol.* **17**, 13-23 (1995).
16. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660-665 (1996).
17. Hénaut, A. & Danchin, A. in *Escherichia coli and Salmonella: cellular and molecular biology* (eds. Neidhardt, F. *et al.*) 2047-2066 (ASM, Washington DC, 1996).
18. Nussinov, R. The universal dinucleotide assymmetry rules in DNA and amino acid codon choice. *Nucleic Acids Res.* **17**, 237-244 (1981).
19. Karlin, S., Burge, C. & Campbell, A. M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **20**, 1363-1370 (1992).
20. Burge, C., Campbell, A. M. & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358-1362 (1992).
21. Kasahara, Y., Nakai, S. & Ogasawara, N. Sequence analysis of the 36-kb region between *gntZ* and *trnY* genes of *Bacillus subtilis* genome. *DNA Res.* **4**, 155-159 (1997).
22. Presecan, E. *et al.* The *Bacillus subtilis* genome from *gerBC* (311°) to *licR* (334°). *Microbiology* **143**, 3313-3328 (1997).
23. Burkholder, P. R. & Giles, N. H. Induced biochemical mutations in *Bacillus subtilis*. *Am. J. Bot.* **33**, 345-348 (1947).

24. Daniels, D. L., Plunkett III, G., Burland, V. & Blattner, F. R. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257**, 771-778 (1992).
25. Wu, L. J. & Errington, J. *Bacillus subtilis* SpoIIIE protein required for DNA segregation during asymmetric cell division. *Science* **264**, 572-575 (1994).
26. Itaya, M. Stability and asymmetric replication of the *Bacillus subtilis* 168 chromosome structure. *J. Bacteriol.* **175**, 741-749 (1993).
27. Billoud, B., Kontic, M. & Viari, A. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res.* **24**, 1395-1403 (1996).
28. Fichant, G. A. & Burks, C. Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* **220**, 659-671 (1991).
29. d'Aubenton Carafa, Y., Brody, E. & Thermes, C. Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* **216**, 835-858 (1990).
30. Stock, J. B., Surette, M. G., Levitt, M. & Park, P. in *Two-Component Signal Transduction* (eds. Hoch, J. A. & Silhavy, T. J.) 25-51 (ASM, Washington DC, 1995).
31. Mizuno, T. Compilation of all genes encoding two-component phophotransfer signal transducers in the genome of *Escherichia coli*. *DNA Res.* **4**, 161-168 (1997).
32. Perego, M., Glaser, P. & Hoch, J. A. Aspartyl-phosphate phosphatases deactivate the response regulator components of the sporulation signal transduction system in *Bacillus subtilis*. *Mol. Microbiol.* **19**, 1151-1157 (1996).
33. Tjalsma, H. *et al. Bacillus subtilis* contains four closely related type I signal peptidases with overlapping substrate specificities: constitutive and temporally controlled expression of different genes. *J. Biol. Chem.* in press (1997).
34. Danchin, A. Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res.* **4**, 9-18 (1997).
35. Suutari, M. & Laakso, S. Unsaturated and branched chain-fatty acids in temperature adaptation of *Bacillus subtilis* and *Bacillus megaterium*. *Biochim. Biophys. Acta* **1126**, 119-124 (1992).
36. Luttinger, A., Hahn, J. & Dubnau, D. Polynucleotide phosphorylase is necessary for competence development in *Bacillus subtilis*. *Mol. Microbiol.* **19**, 343-356 (1996).
37. Landès, C., Hénaut, A. & Risler, J.-L. A comparison of several similarity indices used in the classification of protein sequences: a multivariate analysis. *Nucleic Acids Res.* **20**, 3631-3637 (1992).
38. Glémet, E. & Codani, J.-J. LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci.* **13**, 137-143 (1997).
39. Médigue, C., Moszer, I., Viari, A. & Danchin, A. Analysis of a *Bacillus subtilis* genome fragment using a co-operative computer system prototype. *Gene* **165**, GC37-GC51 (1995).
40. Krogh, S., O'Reilly, M., Nolan, N. & Devine, K. M. The phage-like element PBSX and part of the *skin* element, which are resident at different locations on the *Bacillus subtilis* chromosome, are highly homologous. *Microbiology* **142**, 2031-2040 (1996).
41. Daniel, R. A., Drake, S., Buchanan, C. E., Scholle, R. & Errington, J. The *Bacillus subtilis spoVD* gene encodes a mother-cell-specific penicillin-binding protein required for spore morphogenesis. *J. Mol. Biol.* **235**, 209-220 (1994).
42. Anagnostopoulos, C. & Spizizen, J. Requirements for transformation in *Bacillus subtilis*. *J. Bacteriol.* **81**, 741-746 (1961).
43. Azevedo, V. *et al.* An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artificial chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6047-6051 (1993).
44. Glaser, P. *et al. Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol. Microbiol.* **10**, 371-384 (1993).
45. Ogasawara, N., Nakai, S. & Yoshikawa, H. Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res.* **1**, 1-14 (1994).
46. Sorokin, A. *et al.* A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. *Genome Res.* **6**, 448-453 (1996).
47. Borodovsky, M. & McIninch, J. GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123-133 (1993).
48. Fichant, G. A. & Quentin, Y. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res.* **23**, 2900-2908 (1995).
49. Moszer, I., Glaser, P. & Danchin, A. Subtilist: a relational database for the *Bacillus subtilis* genome. *Microbiology* **141**, 261-268 (1995).

CHAPTER V

Proteome-wide analysis of amino acid frequencies reveals positional biases for specific amino acids and charge- and hydrophobicity gradients between the amino- and carboxy-termini

V.1. Summary

We have analysed fifteen known genomes with respect to positional bias of the amino acids. Amino acid (aa) frequencies, as determined from the entire deduced proteomes (the total of amino acid sequences encoded by a genome), were compared with aa frequencies in the fifteen amino-terminal (N-) and carboxy-terminal (C-) positions of these proteomes. These analyses revealed specific positional biases for many aa's, and these were not restricted to the second (following the methionine) and last (aa corresponding to the codon preceding the stop codon) positions, as has been described by others, but extend well into the first and last fifteen aa positions of the deduced proteomes. Some of the biases appeared to be universal (*i.e.* observed in all proteomes investigated so far), and others are more specific for certain groups of organisms, such as eukaryotes, bacteria, or archaea.

Analysis of the characteristics of the termini with respect to charge and hydrophobicity revealed that all the deduced proteomes investigated displayed similar hydrophobicity- and charge differences between the amino- and carboxy-termini. The average hydrophobicity per aa was found to be lowest in the N-termini, highest in the C-termini, and intermediate in the proteomic average. The calculated charge per aa was highest in the N-termini, lowest in the proteomic average, and intermediate in the C-termini. These findings are not easily reconciled with the assumption that biases are solely the result of mRNA-ribosome interaction effects. Instead, they may reflect the superimposition of other phenomena such as attenuating effects on translation caused by interaction between the ribosome and the nascent polypeptide, conceivably in the post-translational translocation stage of nascent polypeptides through the exit channel(s) of the ribosome. A first test of this hypothesis was performed by analysing subsets of the *Bacillus subtilis* proteome representing class II genes, membrane proteins, long (>500 aa) and short (<100 aa) proteins.

V.2. Introduction

Several studies have been published, primarily based on experimental work in *Escherichia coli*, on the influence of the start codon context on the efficiency of translation. With *lacZ* as a reporter gene, and using immunoprecipitation to measure expression, Looman *et al.* (1987) have investigated the influence of the codon immediately 3' to the AUG initiation codon (*i.e.* position 2 of the aa sequence). Up to a 15-fold difference in expression was observed between the various codons, even between synonymous codons. These differences did not reflect differences in tRNA levels. The usage of codons at the second position was also compared with the overall average codon usage and biases were found for many of them. It was concluded that codon effects at the second position are likely to be due to the interaction of its composing nucleotides with the ribosomal rRNA binding site. Moreover, it was concluded that codon selection at the second position is not based on requirements of the gene product (the protein), but rather determined by factors governing gene regulation at the initiation step of translation. In another study, in which ribosomal selection between two closely positioned AUG start codons was compared, the influence of the 3' flanking nucleotides of the first AUG start codon was measured (Kozak, 1997). This study indicated that the nucleotide composition beyond position +4 (where the A in the AUG start codon is position +1) is generally not affecting recognition by the ribosome.

Several analyses of the stop codon context have been published. These studies focused on the influence of the codons (and thus the encoded aa's) preceding the stop codon on efficiency of termination (Mottagui-Tabar *et al.*, 1994; Bjornsson *et al.*, 1996). Also, Brown *et al.* (1990) have analysed a large number of *E. coli* genes with respect to stop codon context, *i.e.* the three nucleotides before and after the stop codon. Immediately upstream of the stop codon a preference for codons of the form NAR (any of the four bases, adenine, purine) was observed.

In this study, we have analysed aa compositions of the deduced proteome instead of nucleotide context on the gene level. We have extended the analysis to the first and last fifteen amino acids of the deduced proteomes, and carried out these analyses with fifteen entirely known genomes. Biases of aa's at specific positions were observed, and these were shown to extend far beyond the first and last one or two aa's. The effects of these biases on the distribution of charge- and hydrophobicity in these deduced proteomes were investigated. As a first attempt to investigate these findings more closely, we analysed different subsets of the *B. subtilis* proteome representing membrane proteins, proteins encoded by class II genes, long (>500 aa) and short (<100 aa) proteins.

V.3. Methods

Proteome data

Proteomic sequence files in Fasta format were obtained via the www sites of the TIGR institute (TIGR database; http://www.tigr.org/tdb/tdb.html) and the NCBI institute (http://www.ncbi.nlm.nih.gov). A subset of the *Bacillus subtilis* proteome representing membrane proteins was generated using the TopPred2 program (server at:

http://www.biokemi.su.se/~server/toppred2/). Dr. Danchin (Institut Pasteur, Paris) kindly provided us with a list of *B. subtilis* class II genes (genes that are highly expressed during exponential growth; Kunst *et al.*, 1997). Amino acid frequencies were determined from these files with the aid of a program that was written by us for this purpose.

Validation of biases

Relevance of a given deviation from the overall frequency was determined by calculating the *z* value:

$$z = \frac{np^* - np_0}{\sqrt{(np_0(1-p_0))}}$$

In this formula, n is the number of proteins of the proteome, p^* is the observed frequency of an amino acid at a given position, and p_0 is the expected frequency, based on the determined proteomic frequency of the amino acid (Zar, 1996). In this study, with genomes of over 2000 genes, we have taken the absolute value of $z \geq |\mathbf{3,32}|$ as a significant bias in the case of genomes with more than 2000 genes. With smaller genomes, we have set lower critical values for significance. These values are indicated in the Tables V.2A & B.

Charge- and hydrophobicity determinations

To determine the average charge values, we counted positively charged residues (the aa's histidine, lysine & arginine) as +1, and negatively charged ones (aspartate & glutamate) as –1. To determine the average hydrophobicity values, we used the hydrophobicity values of individual amino acids according to Miller *et al.* (1987). These values are listed in Table V.1.

Table V.1. Charge- and hydrophobicity characteristics of amino acids

residue	code	ΔG-Miller (kcal/mol)$^{\Psi}$	charge	residue	code	ΔG-Miller (kcal/mol)$^{\Psi}$	charge
Ala	A	-0.20	0	Leu	L	-0.65	0
Arg	R	1.34	+1	Lys	K	2.00	+1
Asn	N	0.69	0	Met	M	-0.71	0
Asp	D	0.72	-1	Phe	F	-0.67	0
Cys	C	-0.67	0	Pro	P	-0.44	0
Gln	Q	0.74	0	Ser	S	0.34	0
Glu	E	1.09	-1	Thr	T	0.26	0
Gly	G	-0.06	0	Trp	W	-0.45	0
His	H	-0.04	+1	Tyr	Y	0.22	0
Ile	I	-0.74	0	Val	V	-0.61	0

$^{\Psi}$ ΔG-Miller: empirical hydrophobicity scale of Miller *et al*. (1997). Negative values represent hydrophobic residues; positive values are hydrophilic.

V.4. Results and discussion

Positional analysis of amino acids

We have analysed the proteomes of *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Caenorhabditis elegans*, *Saccharomyces*

cerevisiae, *Synechocystis* sp. strain PCC6803, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *B. subtilis*, *E. coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Treponema pallidum*, *Borrelia burgdorferi*, and *Aquifex aeolicus*. Overall frequencies of the aa's were determined for each proteome, and these were compared with the frequencies in the fifteen N-

Table V.2A. Positional bias of amino acids in the N-termini of known proteomes

AA	Archaea			Eukarya		Bacteria									
						Cy	Low G+C G^+			Gamma subdiv			Spiroch		Aqui
	Af	Mt^3	Mj^3	Ce	Sc	Sp	$Mg^{2,5}$	$Mp^{2,5}$	Bs	Ec	Hi	Hp^3	$Tp^{2,5}$	$Bb^{2,5}$	Aa^3
A	3, 5↓	2-5, 6, 8↓	2-5↓	2↑; 5, 6, 12-14↓	2↑; 11-13↓	2↑; 3, 5, 6↓	2↑; 3, 5↓	3, 8↓	3-7↓	3-6↓	3-6↓	3, 5↓	5, 6↓	3, 5↓	3, 4, 8↓
C		13↑	11↑	3↓	2↓	2↓	12↑	7, 12, 15↑	11, 15↑	2, 3↓	13↑	14, 15↑	5, 6, 7, 10-12, 14, 15↑		
D	4↓	6, 15↓	6, 7↓	5-15↓	3-12, 15↓	4, 7↓	2, 6↓	4↓	2-4, 6, 7, 13-15↓	2, 4, 6, 12↓	2, 4, 9, 13, 15↓	5↓	7, 13↓	5, 6↓	3, 13↓
E		2, 4, 9, 12↓	4↓	2, 5-15↓	2, 6-15↓	3-13↓	3, 7↓	4, 5, 6, 7↓	4, 6-9, 14↓	2, 4-9, 11↓	2, 5, 6, 9↓	6, 10, 11↓	7, 8, 12, 13↓	4, 9, 12, 13, 15↓	4, 6, 7↓
F	2↓	6↑	3↑	2↓; 4-15↑	2↓	4, 5, 6↑	2↓	3↑	2, 3↓	2↓	4, 5, 7, 8↑	3, 4, 10, 11↑	2↓	5, 6, 8, 14↑	4, 6, 11, 12, 15↑
G	2-6↓	2, 4, 5↓	3-6↓	3-11↓	4-8↓	2-11↓	3, 4, 6↓	3, 4, 5, 7↓;15↑	2-8↓	2-8↓	2-8↓	2-6↓		2, 3, 5, 6, 8↓	3-5, 6↓
H				9-11↓	2↓				2↓						
I	5, 6, 9↑	2, 3, 5-8↑	2-8, 10↑	2↓; 5-8, 10-13↑	2↓	3, 5↑	9↑	2↓; 5, 7, 8↑	2↓; 4-10↑	4-8, 13↑	3-8↑	4, 5↑		7, 15↑	5-7, 8↑
K	2-4↑	2-12↑	2, 3, 6, 8↑	5↑; 8, 11-13, 15↓	2↓; 5, 7↑	2-4, 6, 7↑	2-6, 9↑	2-7↑	2-7↑	2-6↑	2-7↑	2-4↑	2, 3↑	2-5, 7, 8↑	2-4↑
L	2↓; 4, 6, 7, 11, 12, 14↑	2↓	2, 10↓	2↓; 3-15↑	2, 3↓	2, 3↓; 5, 8-11↑	7, 12↑	2↓; 7, 11↑	2, 3↓; 8-14↑	2, 3↓; 6, 8-15↑	2↓; 5, 6, 10, 11↑	6-10, 13, 14↑	2, 3↓	6, 9-12, 13↑	2↓; 5-11, 13↑
M	4↑; 10↓	2, 4, 6↑	2-6, 11↑	2-15↓	4-15↓				2↓	7, 8↓					8, 11↓
N		2↑	8↓	2↑; 9, 10, 12-15↓	14↓	2-4↑	4↑		2-4↑	2-4↑	2, 3↑		2↑	2↑	5, 7, 8↓
P		2, 4↓	3, 4, 13↓			4-9↑			2-7, 9, 14↓	3↓	3↓	3↓		3, 6, 10↓	2, 3↓
Q				2↓; 7-10, 12, 15↑	2↓; 4, 5↑	2↓			2↓; 3-5↑	3↑	4↑	2↑			
R	2, 3↑		2, 11↑	2, 3↑	2↓; 3-11, 15↑	3↓	3↑	2, 12↑	3-5↑			2↑	2, 3, 4, 5, 6, 10↑	2, 3, 4↑	2, 3, 4↑
S			5↓	2-8, 11, 14, 15↑	2-4, 6, 13↑	2-8, 10-12, 14↑		7↓; 11↑	2↑	2, 4↑	2, 8↑		2↑	5↓	2, 15↑; 4↓
T	2↓		2↓	2, 3↑	2, 5, 6↑	2-4, 12↑		11↓	2, 3↑	2, 3, 5, 6↑	2, 3, 5↑	2↓			
V	3-5↑		2↑	2, 3↓; 13↑		3, 4↓			2, 3↓	2, 3↓	2, 3, 5↓	2↓	2↓; 12↑		2, 9↓
W			11↑	3, 11↓	2↓	2↓			2↓	2↓			2↓		
Y	2↓	2↓	2↓	2, 3, 6, 11↓	2,6↓	2↓	2↓	2↓	2↓	2↓	2↓	2↓			2↓

See Table V.2B for legend.

Table V.2B. Positional bias of amino acids in the C-termini of known proteomes

	Archaea			**Eukarya**		**Bacteria**									
						Cy	**Low G+C G$^+$**			**Gamma subdiv**			**Spiroch**		**Aqui**
AA	**Af**	**Mt^3**	**Mj^3**	**Ce**	**Sc**	**Sp**	**Mg^{2,5}**	**Mp^{2,5}**	**Bs**	**Ec**	**Hi**	**Hp^3**	**Tp^{2,5}**	**Bb^{2,5}**	**Aa^3**
A	1-4↓	1-4, 6↓	1-4↓	1-3, 5, 6↓	2↓				1-6↓		3, 6↓	3, 5↓	2↓	1-4, 8, 10↓	2-4, 7, 14↓
C		12↑		1, 4↑					14↑				7↑		
D	2-5↓	2-4↓	2↓	1, 2, 4↓		6↓			2, 3, 5, 6↓	4↓		3↓	1↓	6↓	2, 3, 5↓
E	4, 8, 10, 11, 14↑	1, 6↑	1, 4, 11, 13↑	2, 3↓	8-10, 12↑	7, 12, 13↑	4↑		1, 5-8, 11, 13↑	1, 3↑	1↑	3, 11↑	1↑	7, 8↑	4-8, 10, 12↑
F				1-3, 5↑	2, 3↑	1↑				2↓		1↑			
G		3↑; 7↓	4, 6↓	1-3, 7, 8↓	1↓	5, 8, 9, 11-13 ↓	2, 10↓	7, 9, 11↓	1, 3, 6, 8-13↓	3, 5↓	1, 5, 7↓	10, 11↓		3, 5, 11↓	6, 11, 12↓
H	1↓	12↓		1↑					2↑	1↑	1↑	1↑			
I	2↓	8↑			4↓	1, 2↓			1, 2, 4↓	1, 2, 4↓		1↓			1-3↓
K	2, 4, 5, 7, 8, 13-15↑	1-7, 9-12↑	1-8, 11, 15↑	1-12, 14↑	1-9, 11-15↑	1-3, 5-7, 10↑	1-3, 5-7, 12↑	1, 2, 4-6, 8, 10↑	1-13↑	1-6↑	1-7, 14↑	1-5, 7, 8↑	2, 6↑	1-7, 15↑	2-10, 13, 14↑
L	3, 5, 6↑		3, 4↑	4, 5, 8, 9, 12-15↓	4↓	1↓	1↓	1↓	1, 2↓	1, 4↓				1↓; 5↑	
M	1↓	1↓	1↓	1, 3, 4, 8, 9, 11, 15↓		1↓		4↓	2↓	1, 3, 4, 12↓	1↓		2, 11↓	15↓	1↓
N	1↓	1↓		1, 2, 5, 8, 9↑	1↑; 10↓	1, 2↑	1, 2↑	1↑	2↑	10↓	2↑			1, 2↑	
P	2, 7, 10↓		1-3, 5-7↓	1, 2↓	1, 2↓		1, 2↓	1↓	1, 2, 5, 7↓	1↓	1, 2↓	2↓	5↓	2↓	2, 6, 9, 10↓
Q	3↑		1↑	13↑		1↓	4↑		6, 9↑	1↑					
R	1-7, 9-11, 14, 15↑	1-7, 12↑	1-4, 7↑	2, 3, 5-11↑	2, 4, 5, 7-10↑	1, 6, 9↑	8, 12↑	2, 6, 8, 10, 13, 15↑	1-7, 10↑	1-4, 7, 8, 12, 15↑		1, 2, 13, 14↑	1, 2, 7, 8, 10, 11, 13, 14↑	3↑	1, 2, 4, 14↑
S			6, 11↓	1↓; 3, 4↑	1, 9, 11, 12↓	1-5↑	1↑	6, 9↓	1, 2↑				2↑	7↓	1, 2↑
T	2-4, 7↓	1, 3↓	1, 3, 4, 9↓	1, 2, 12↓; 3↑	1, 3, 7, 8, 14↓	1↓	1↓	1, 2↓	1, 5↓	1↓	1↓	1↓		1, 2, 4↓	5↓
V		1, 2↓	1, 2↓	2↓	2↓	2↓	1↓	1↓		1↓	1↓				
W	1-4, 10↑	2, 7↑	1↑		1, 10↑	8↑	4↑						1, 4↑		3↑
Y	1↓		2↓	1↑		2↓	1, 11↑				3↑	5, 8↑	2↓		

In the first column, the aa's are listed in the one-letter code (see Table V.1.). In subsequent columns, the observed biases of these aa's in the respective proteomes are listed. The numbers represent the positions; position 1 in the N-terminus corresponds to the first aa (methionine), and position 1 in the C-terminus is the last amino acid (encoded by the codon preceding the stop codon). Arrows pointing up indicate over-representation at that position; arrows pointing down indicate under-representation. Abbreviations of organism names (and the number of proteins analysed): **Archaea** *Archaeoglobus fulgidus* (Af; 2409); *Methanobacterium thermoautotrophicum* (Mt; 1872); *Methanococcus jannaschii* (Mj; 1715); **Eukarya** *Caenorhabditis elegans* (Ce; 13450); *Saccharomyces cerevisiae* (Sc; 6187); **Bacteria/Cyanobacteria** (Cy) *Synechocystis* sp. strain PCC6803 (Sp; 3168); **Bacteria/Low G+C Gram-positive bacteria** (Low G+C G$^+$); *Mycoplasma genitalium* (Mg; 467); *Mycoplasma pneumoniae* (Mp; 677); *Bacillus subtilis* (Bs; 4100); **Bacteria/Gamma subdivision** (Gamma subdiv) *Escherichia coli* (Ec; 2438); *Haemophilus influenzae* (Hi; 1713); *Helicobacter pylori* (Hp; 1577); **Bacteria/Spirochaetales** (Spiroch) *Treponema pallidum* (Tp; 1031); *Borrelia burgdorferi* (Bb; 850); **Bacteria/Aquificales** (Aqui) *Aquifex aeolicus* (Aa; 1522). In superscript with the organism name, the cut-off value of the *z* score is indicated when a value other than 3,32 was taken to correct for smaller genome sizes.

terminal and C-terminal aa's of the same proteome. In this work, the N-terminal initial methionine is defined as position 1; the C-terminal aa corresponds to position 1'. Thus, we have analysed the proteomic positions 1 to 15 and 1' to 15' of the organisms listed above. By calculating the corresponding *z* values of the deviation from the overall frequencies, an indication of the relevance of the observed bias at each position was obtained. The results of these analyses are summarised in Tables V.2A (N-termini) and V.2B (C-termini). In the analyses of the N-termini, biases at the first position were omitted, as methionine almost invariably occupies this position. The aa frequency data and their *z*-values can be obtained at URL: [www.biol.rug.nl/molgen/noback.html].

Biases that are (almost) universal

In the N-termini, biases that are (almost) universal were observed for alanine, glycine, lysine, leucine, and tyrosine. Alanine is under-represented in all proteomes at positions three and/or five, with the exception of *S. cerevisiae*. Glycine is generally under-represented in the N-termini, at least at several positions between 2 to 8, except in *T. pallidum*. The under-representation of glycine is most pronounced in the proteomes of *Synechocystis* and *C. elegans*. Except for the eukaryal ones, all proteomes have an over-representation of lysine in the N-terminus, at least at positions 2 and 3, and generally extending up to position 12. Leucine is always under-represented at position 2, except in *M. genitalium*, *H. pylori*, and *B. burgdorferi*. Tyrosine is under-represented at position 2 in all proteomes, except in those of *T. pallidum* and *B. burgdorferi*. These two proteomes do in fact have a decreased tyrosine frequency at position 2, but the *z* score of this bias was below the cut-off value.

In the C-termini of the proteomes, three (almost) universal biases were observed, for lysine, arginine, and threonine. Lysine is over-represented to a greater or lesser extent in all proteomes, but this bias is not restricted to fixed positions. The same applies to arginine. The latter aa is over-represented in the C-termini of all proteomes, with the exception of *H. influenzae*. Threonine is under-represented at least at position 1' in all proteomes, except for *A. aeolicus*, *T. pallidum*, and *A. fulgidus*.

Group-specific biases

In the analysis of the N-termini, presently restricted to only two eukaryal proteomes, the frequency of alanine is decreased at several more distal positions: 12-14 (*C. elegans*) or 11-13 (*S. cerevisiae*). This aa is over-represented at position 2 in the eukaryal proteomes and those of *Synechocystis* and *M. genitalium*. The negatively charged aa's aspartate and glutamate are strongly under-represented in most of the N-terminal positions of the eukaryal proteomes. Methionine is under-represented in almost the entire N-terminus of the eukaryal proteomes, at least between positions four and fifteen. Albeit to a lesser extent, the archaeal proteomes have the reverse of this bias, with an over-representation of methionine between positions 2 and 6.

In the C-termini, the archaeal proteomes have an under-representation of asparagine at position 1' (not indicated as such for *M. jannaschii* in Table V.2B, since the *z* value was just below the cut-off value). The archaeal proteomes also have an under-representation of alanine

at positions 1' to 4', which also applies to *B. subtilis* and *B. burgdorferi*. In the two eukaryal proteomes analysed, serine is under-represented at position 1'. Histidine is over-represented at position 1' in bacteria of the Gamma subdivision, also in *C. elegans*.

Comparison with data from the literature

Looman *et al.* (1987) have analysed effects of the codon at the second position (coding for aa 2) on translation efficiency in *E. coli*, and the frequency of codons at this position in comparison with average genomic frequencies. In these analyses, poor expression correlated with the codons UUC (major codon for phenylalanine; 67%), UCA (minor codon for serine; 6%), and CUG (major codon for leucine; 74%) at the second position. However, contrary to our observations, under-representation of phenylalanine at the second position was not found. Where we observed a proteomic frequency (p.f.) of phenylalanine of 3.86% and 2.79% at the second position (2^{nd} pos.), they reported a p.f. of 3% and 3.8% at the second position. Since Looman *et al.* used old codon frequency data from 1984, we consider it likely that the under-representation of phenylalanine we observed at the second position is more in line with the observed expression data. The observation of Looman and co-workers that leucine had a p.f. of 8.4% and a frequency of 2% at position 2, is only slightly reflected in our data: we observed the frequencies 10.6% and 7.2%, respectively. For serine, our data (over-represented at the second position: p.f. 5.6%; 2^{nd} pos. 14.3%) are in accordance with those of Looman *et al.* (p.f.: 5.1%; 2^{nd} pos. 15.9%), although we observed an over-representation at position 4 as well (8.3%). Interestingly, Looman and co-workers found the highest expression efficiencies in their *lacZ* reporter system with the major codon for lysine in the second position, being encoded by the AAA trinucleotide (which is used in 72% of the codons for lysine). They observed a high over-representation of this codon (p.f. 5.2%; 2^{nd} pos. 13.9%), and the corresponding aa (p.f. 7.2%; 2^{nd} pos. 15.3%). This is in agreement with what we observed for all proteomes (except the eukaryal ones) and, moreover, the lysine over-representation was not restricted to the second position. Therefore, it can be questioned whether the nucleotide context effects on translation may not be restricted to the codon immediately downstream of the start codon, but could possibly extend up to 20 nucleotides downstream of the start codon. So far, this idea has not been substantiated in the literature. Kozak (1997) observed that nucleotide effects on (eukaryal) expression (measured through ribosomal selection of the start codon versus a second, constant, start codon) do not extend beyond position +5 (the second nucleotide of the codon following ATG). In this paper, it was reported that expression was improved when a G residue was present at the first position of the second codon, except when it was placed in the codon GUA (valine). Especially, the codons GCG, GCU, GCC, GCA (all for alanine), GAU (for asparagine), and GGA (for glycine) improved expression. For alanine, this is reflected in our data on the N-terminus of the eukaryal proteomes. Alanine is over-represented at position 2 in *C. elegans* (p.f. 6.2%; 2^{nd} pos. 8.6%) and *S. cerevisiae* (p.f. 5.5%; 2^{nd} pos. 8.2%). In contrast, neither aspartate nor glycine is over-represented at position 2 in *C. elegans* and *S. cerevisiae*. Valine is under-represented in *C. elegans* (p.f. 6.2%; 2^{nd} pos. 4.4%), but not in yeast (p.f. 5.6%; 2^{nd} pos. 6.4%). GUA is used in 16% of the codons for valine in *C. elegans* and in 21% of these codons in *S. cerevisiae*.

With respect to the C-terminus, Brown *et al.* (1990) have analysed the nucleotide context of the stop codon region of 862 *E. coli* genes. Preceding the UAA stop codon, a preference for codons of the form NAR (any of the four bases, adenine, purine) was observed and, in particular, those that code for glutamine or the basic aa's (histidine, lysine, arginine). In contrast, codons for threonine or branched nonpolar amino acids were under-represented. In our analyses, glutamine (p.f. 4.4%; 1' pos. 6.3%), histidine (p.f. 2.3%; 1' pos. 4.0%), lysine (p.f. 4.5%; 1' pos. 11.2%), and arginine (p.f. 5.7%; 1' pos. 8.7%) were found to be over-represented at position 1', while threonine was found to be under-represented (p.f. 5.3%; 1' pos. 1.1%). This is in accordance with the findings of Brown and co-workers (1990) and, moreover, the over-representations of lysine and arginine are not restricted to position 1', but extend to position 6' (lysine) and 15' (arginine) in *E. coli*.

Analysis of subsets of proteins in the *Bacillus subtilis* proteome

We have investigated whether aa biases at certain positions in the proteomes could be related to certain classes of proteins. To this end, we have analysed subsets of proteins from the *B. subtilis* proteome representing class II gene products (highly expressed during exponential growth), (putative) membrane proteins, long (>500 aa) and short (<100 aa) proteins. The data of these analyses are summarised in Table V.3.

Analysis of the different *B. subtilis* proteomic subsets revealed some specific features, in particular in the class II proteins subset. In the N-terminus of the class II subset, alanine is highly over-represented at position 2 (p.f. 8.6%; 2^{nd} pos. 22.5%), while this is not the case in the proteomic average (p.f. 7.7%; 2^{nd} pos. 7.1%). Serine is also much more over-represented at position 2 in the class II subset (p.f. 5.4%; 2^{nd} pos. 16.3%) than in the proteome (p.f. 6.3%; 2^{nd} pos. 9.8%). This also applies to lysine in the class II subset (p.f. 7.9%; 2^{nd} pos. 15.3%, 3^{rd} pos. 23.4%) than in the proteome (p.f. 7.1%; 2^{nd} pos. 19.3%, 3^{rd} pos. 17.5%). Interestingly, in the subset of long proteins, similar effects were observed with respect to serine (p.f. 6.4%; 2^{nd} pos. 12.5%) and lysine (p.f. 6.9%; 2^{nd} pos. 18.3%, 3^{rd} pos. 24.1%), but not with alanine (p.f. 7.9%; 2^{nd} pos. 10.0%). In contrast, the subset representing short proteins shows only minor over-representation of serine (p.f. 6.2%; 2^{nd} pos. 9.9%) and lysine (p.f. 8.6%; 2^{nd} pos. 12.1%, 3^{rd} pos. 14.7%). Except for the class II subset of proteins, asparagine is over-represented at position 2 in all subsets, including the total proteome.

In the C-terminus, lysine is more over-represented at position 1', but not at subsequent positions, in the class II subset (p.f. 7.9%; 1^{st} pos. 23.0%, 2^{nd} pos. 9.6%), when compared to the proteome (p.f. 7.1%; 1^{st} pos. 15.2%, 2^{nd} pos. 11.2%). In the subset of short proteins, this over-representation of lysine is, as in the N-terminus, less pronounced (p.f. 8.6%; 1^{st} pos. 14.2%, 2^{nd} pos. 9.9%). Threonine is more biased at position 1' in the class II subset of proteins (p.f. 5.7%; 1^{st} pos. 1.0%) than in the proteome (p.f. 5.4%; 1^{st} pos. 2.5%). The bias of this aa is intermediate in the membrane protein subset (p.f. 5.5%; 1^{st} pos. 1.8%), and about equal to the proteomic bias in the subset of long proteins (p.f. 5.6%; 1^{st} pos. 2.1%) and the subset of short proteins (p.f. 4.8%; 1^{st} pos. 2.2%).

We did not observe stronger biases at the N-terminal positions 4 to 10 for the hydrophobic amino acids in the membrane protein subset, as compared to the total proteome.

This indicates that the leucine and isoleucine biases at these positions in the proteome are not simply the result of the large fraction of membrane proteins in the genome.

Table V. 3. Positional biases of aa's in subsets of the *B. subtilis* proteome

	N-terminus						C-terminus				
AA	proteome (4100)	class II (209)	mem-brane (221)	long (241)	short (232)	**AA**	proteome (4100)	class II (209)	mem-brane (221)	long (241)	short (232)
A	3-7↓	2↑↑	5↓	3↓	5↓	**A**	1-6↓	3↓		1↓	3, 4↓
C	11, 15↑					**C**	14↑				
D	2-4, 6, 7, 13-15↓	2↓		7↓	9, 15↓	**D**	2, 3, 5, 6↓	3, 4↓		2↓	5↓
E	4, 6-9, 14↓	2, 6, 13↓		6, 11↓	4↓	**E**	1, 5-8, 11, 13↑	9↑	2↑	9,10↑	1,7↑
F	2, 3↓	2↓				**F**					
G	2-8↓	2, 3, 5↓	2-7, 10↓	5, 8↓	2, 4, 5↓	**G**	1, 3, 6, 8-13↓	3↓		3↓	
H	2↓	2↓	2↓	2↓		**H**	2↑	5↓	2↑		8↑
I	2↓; 4-10↑	2↓	10↑	2↓; 7↑	7↑	**I**	1, 2, 4↓	1, 4, 7, 14↓	4↓	5↓	7↓
K	2-7↑	3↑↑; 2, 4-7↑	2, 3, 5, 9↑	2↑, 3↑↑	2, 3↑	**K**	1-13↑	1↑↑; 4↑	2, 3↑	1, 2, 4, 6↑	1, 4↑
L	2, 3↓; 8-14↑	2↓		3↓		**L**	1, 2↓	1, 4, 6↓	4↓		1↓
M	2↓	2↓				**M**	2↓		2↓		
N	2-4↑		2-4, 6↑	2↑	2↑	**N**	2↑	2, 3↑		1, 8↑	7, 8↑
P	2-7, 9, 14↓	3↓		3, 4↓		**P**	1, 2, 5, 7↓	1, 2, 5↓			
Q	2↓; 3-5↑	2↓; 3, 10↑	6↑	3↑	2↓	**Q**	6, 9↑	2, 14↑	1↑		
R	3-5↑	3↑	2-5↑	4,5↑		**R**	1-7, 10↑	9, 12↑	10, 11↑	1↑	2↑
S	2↑	2↑↑	2↑	2↑↑	2↑	**S**	1, 2↑	2↑	3, 5↓		2↑
T	2, 3↑	7, 9↑	3, 7↑	2↑	7↓	**T**	1, 5↓	1↓↓	1↓	2↓	1↓
V	2, 3↓	2, 3↓	2, 3↓	2, 3↓	2↓	**V**		3, 4, 6↑	6↓		1↓
W	2↓		2↓	2↓		**W**		8↑	11↓		
Y	2↓	2↓	3↓	2↓		**Y**				2↓	

In the first and seventh column, the aa's are listed in their one-letter code. In subsequent columns, biases of these aa's at certain positions in the *B. subtilis* proteome and its respective subsets are listed. The numbers of protein sequences analysed are indicated between parenthesis. Class II: genes that are highly expressed under exponential growth conditions; membrane: membrane proteins according to the Toppred2 prediction program; long: proteins of > 500 aa; short: proteins of < 100 aa. The numbers represent the positions; position 1 in the N-terminus corresponds to the first aa (methionine), and position 1 in the C-terminus to the last aa (encoded by the codon preceding the stop codon). Arrows pointing up indicate over-representation at that position; arrows pointing down indicate under-representation.

Analysis of charge and hydrophobicity distribution

Since all proteomes analysed here displayed many species-specific biases for aa's, we wondered whether in various organisms different variations might have evolved that result in the same properties, just as similar protein architectures can be obtained with dissimilar aa sequences. To address this question we analysed, using the aa frequency data of the N-termini, the C-termini and those of the whole proteome, the charge- and hydrophobicity characteristics of both termini, and compared those with the proteomic average. From these analyses, the results of which are presented in Fig. V.1, some general features of protein characteristics seem to emerge at the proteomic level. The calculated average charge (per aa;

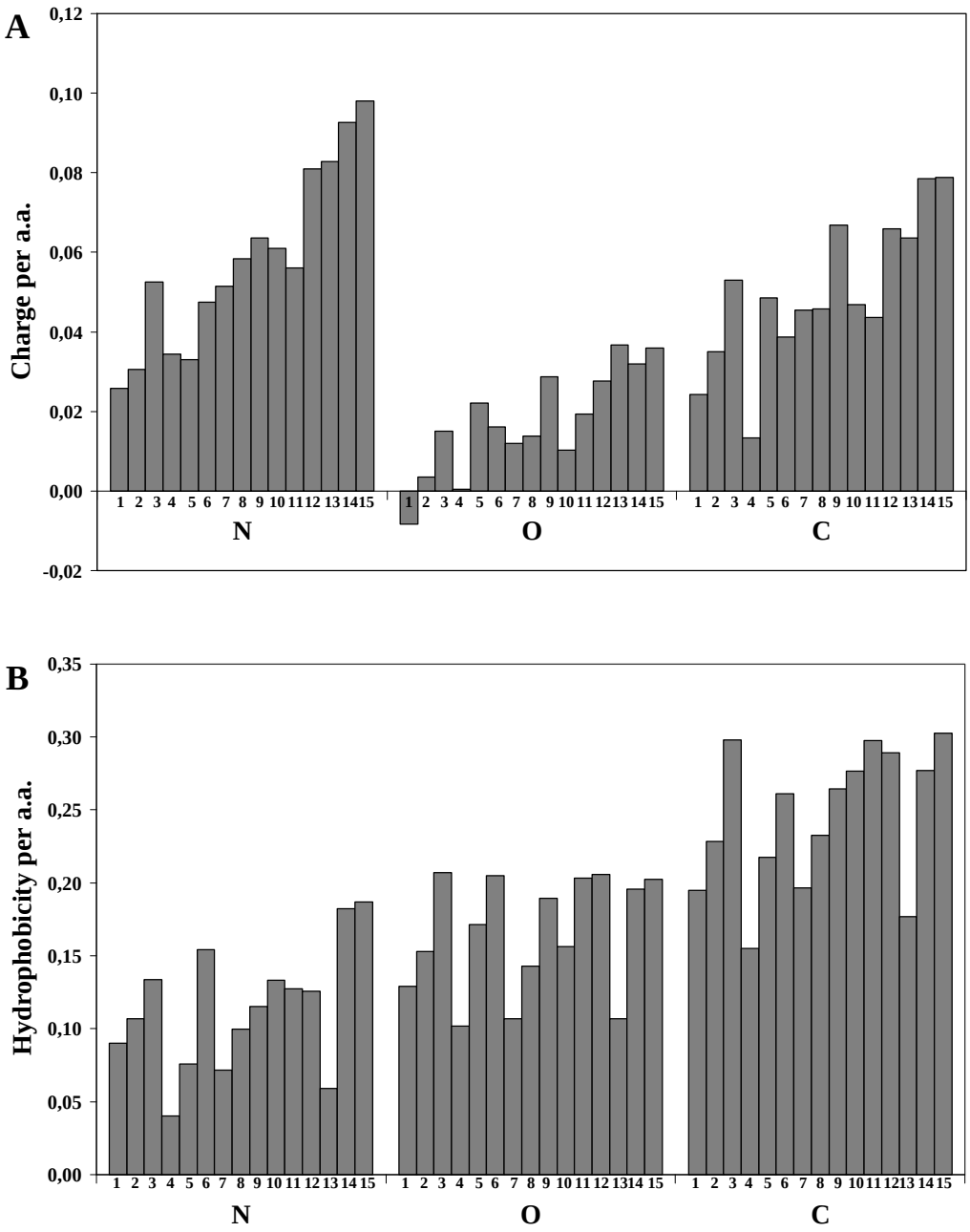


Fig.V.1. Charge- and hydrophobicity values per aa of whole proteomes (O), and the corresponding values for their N- and C-terminal fifteen aa's. Panel **A**: average charge per aa. Panel **B**: average hydrophobicity per aa (kcal/mol; the positive values represent hydrophilicity). Organisms are grouped with respect to archaeal (1-3), cyanobacterial (4), eukaryal (5-6), and bacterial origin (7-15). 1: *M. thermoautotrophicum*; 2: *A. fulgidus*; 3: *M. jannaschii*; 4: *Synechocystis sp*; 5: *C. elegans*; 6: *S. cerevisiae*; 7: *E. coli*; 8: *H. influenzae*; 9: *H. pylori*; 10: *B. subtilis*; 11: *A. aeolicus*; 12: *B. burgdorferii*; 13: *T. pallidum*; 14: *M. pneumoniae*; 15: *M. genitalium*.

Fig. V.1A) shows similar differences between the N-terminus, the C-terminus, and the average proteome in all investigated organisms. The N-terminal aa's have the highest positive charge, the average proteome the lowest, and the C-terminal aa's are intermediate to these. Only one proteome has a negative average charge (per aa): *M. thermoautotrophicum*.

With respect to the hydrophobicities, we also observed a kind of gradient (Fig. V.1B). The hydrophobicity is lowest in the N-terminal aa's, highest in the C-terminal aa's, and intermediate in the proteomic average. The observation that the termini of proteomes have the highest average charge per aa as compared to the overall average, is consistent with the fact that the termini of proteins are generally located at the outside of the protein in the three-dimensional structure, and thus are solvent-exposed. This is also reflected in the hydrophobicity characteristics of the C-termini of the proteomes, which are most hydrophilic, but not in the N-termini. Since the charge- and hydrophobicity values of the N- and C-termini seem to be interrelated, we plotted the average charge- and hydrophobicity values of the N-terminal aa's against the average charge- and hydrophobicity values of the C-terminal aa's. Fig. V.2 shows these relationships. It should be noted here that charge and hydrophobicity are not entirely independent parameters; when a charge is present, this influences the hydophobicity characteristics.

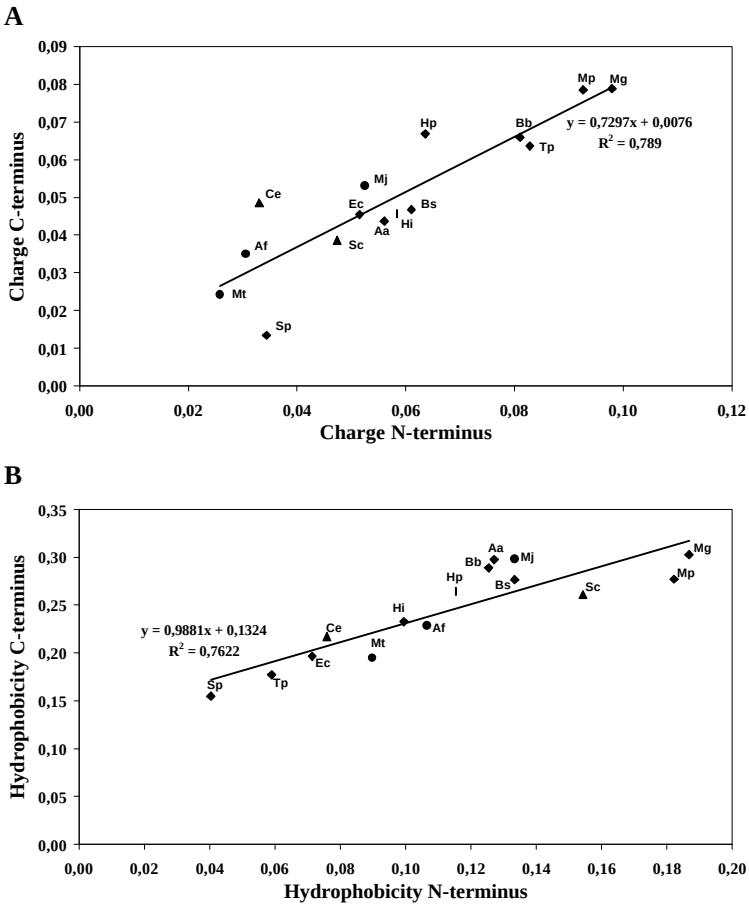


Fig. V. 2. Analysis of whole proteomes with respect to (**A**) the relationship between the charge of the N-terminus and the C-terminus and (**B**) the relationship between the hydrophobicity of the N-terminus and the C-terminus (per aa). Each point represents a proteome, and the respective organisms are indicated in the same two-letter code as was used in Table V.1. Archaea are indicated with a circle, eukarya with a triangle, and bacteria with a square. Regression lines, with the respective formula and R^2 value, are shown in the figures.

All proteomes investigated thus far displayed similar differences between the proteomic average (O), the N-terminus, and the C-terminus with respect to the charge- and hydrophobicity distribution. This was also found to be the case with different subsets of the *B. subtilis* proteome that were analysed (see below). We feel that these findings are likely to reflect a phenomenon with biological relevance. The only shared feature of all proteins, being

cytoplasmatic, membrane-associated, or secreted, is that they are ribosomally synthesised. Therefore, the observed characteristics could, for instance, be related to the translation process, and in particular to the architecture of the translation machinery. The three-dimensional structure and functional characteristics of the (*E. coli*) ribosome, in particular the decoding centre, have been elucidated to some extent in recent years (Frank *et al.*, 1991; Frank *et al.*, 1995; Nierhaus *et al.*, 1995; Easterwood & Harvey, 1995; Nierhaus *et al.*, 1998). However, the fate of the nascent polypeptide and its interaction with the ribosome are not well understood. Frank and co-workers (1995) have suggested possible exit route(s) for nascent polypeptides through the 50S subunit of the *E. coli* ribosome. Starting at the decoding centre, a tunnel (T) runs through the 50S subunit, splitting up into two branches (T1 and T2) that exit the ribosome at 65 Å from each other. It was postulated that one exit (E1) would be associated with the cytoplasmatic membrane, whereas the other exit (E2) would open directly into the cytoplasm. T starts as a narrow tunnel (10-15 Å in diameter; which is sufficiently wide to allow a polypeptide in α-helical conformation to pass), and it widens (25-30 Å) before splitting up into T1 and T2. A cavity (C) is associated with tunnel T2 close to its exit point E2, and this cavity possibly plays a role in chaperone-assisted folding of the nascent polypeptide. Conceivably, T1 is the exit tunnel for exported proteins and membrane proteins, while cytoplasmatic proteins are likely to be transported through T2 (Frank *et al.*, 1995). Since the first part of the tunnel T is just wide enough to accommodate the nascent polypeptide, one can easily imagine some (hydrophobic- and/or electrostatic-) interaction taking place between the tunnel and nascent polypeptide, thus influencing the rate of protein release from the ribosome. Unfortunately, the characteristics of the exit tunnels with respect to charge and hydrophobicity are not known.

The specific biases of the *B. subtilis* proteomic subsets of proteins (Table V.3), and the resulting charge- and hydrophobicity characteristics they entail, could reflect such a phenomenon. The charge- and hydrophobicity characteristics of the *B. subtilis* proteome and its subsets of class II gene products and membrane proteins are presented in Fig. V.3. In addition to the specific aa biases, the class II subset is N-terminally more hydrophilic than the

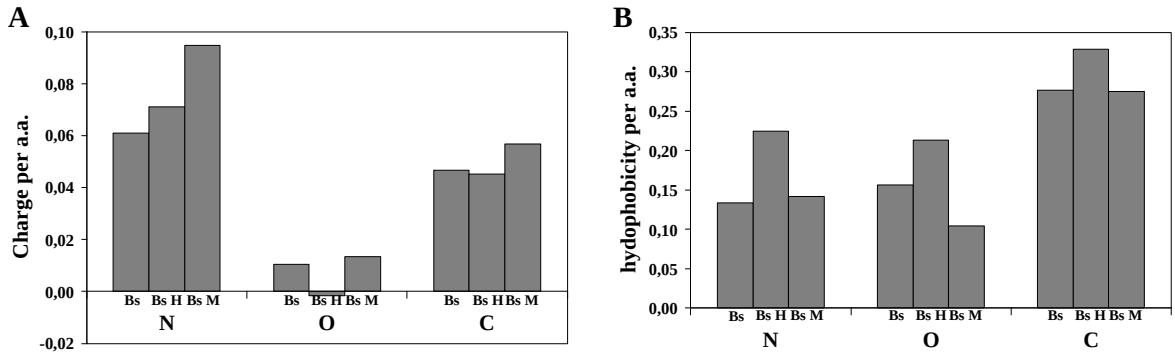


Fig. V.3. Charge- and hydrophobicity characteristics of the *B. subtilis* proteome (Bs) and its subsets of class II genes (Bs H) and membrane proteins (Bs M). **A:** Average charge per aa in the N-terminus (N), overall (O), and in the C-terminus (C). **B:** Average hydrophobicity (kcal/mol) in the N-terminus (N), overall (O), and in the C-terminus (C).

proteomic average (0.224 kcal/mol and 0.133 kcal/mol, respectively).

When the observed differences in charge- and hydrophobicity characteristics between species and groups of species (Fig V.1 & V.2) are indeed related to the interaction between nascent polypeptides and the ribosomal channel(s), these differences should be reflected in the specific properties of the ribosomal channel from the respective organisms.

With respect to the two-tunnel hypothesis for polypeptide exit from the ribosome, a signal would be necessary that directs the nascent polypeptide to the correct tunnel and respective exit. The N-terminal sequence of the polypeptide would be a likely candidate to harbour such a signal. In the charge- and hydrophobicity analyses of *B. subtilis* proteomic subsets, we observed some deviations in the subset of membrane proteins that may constitute such a signal. The N-terminus of the membrane protein subset is more positively charged than that of the total proteome (0.095 and 0.061, respectively). This is not really surprising, since positively charged residues often precede membrane-spanning domains. However, this feature could also be used as a signal to direct nascent polypeptides to the correct exit from the ribosome.

References

Bjornsson, A., Mottagui-Tabar, S., & Isaksson, L. A. (1996). Structure of the C-terminal end of the nascent peptide influences translation termination. *EMBO J.* **15**, 1696-1704.

Brown, C. M., Stockwell, P. A., Trotman, C. M., & Tate, W. P. (1990). The signal for the termination of protein synthesis in procaryotes. *Nucleic Acids Res* **18**, 2079-2086.

Easterwood, T. R. & Harvey, S. C. (1995). Modeling the structure of the ribosome. *Biochem.Cell Biol.* **73**, 751-756.

Frank, J., Penczek, P., Grassucci, R., & Srivastava, S. (1991). Three-dimensional reconstruction of the 70S *Escherichia coli* ribosome on ice: The distribution of ribosomal RNA. *J.Cell.Biol.* **115**, 597-605.

Frank, J., Verschoor, A., Li, Y., Zhu, J., Lata, R. K., Radermacher, M., Penczek, P., Grassucci, R., Agrawal, R. K., & Srivastava, S. (1995). A model of the translational apparatus based on a three-dimensional reconstruction of the *Escherichia coli* ribosome. *Biochem.Cell Biol.* **73**, 757-765.

Kozak, M. (1997). Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.* **16**, 2482-2492.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Düsterhöft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S.,

Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, H., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., & Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Looman, A. C., Bodlaender, J., Comstock, L. J., Eaton, D., Jhurani, P., de Boer, H. A., & van Knippenberg, P. H. (1987). Influence of the codon following the AUG initiation codon on the expression of a modified *lacZ* gene in *Escherichia coli*. *EMBO J.* **6**, 2489-2492.

Miller, S., Janin, J., Lesk, A. M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *J.Mol.Biol.* **196**, 641-656.

Mottagui-Tabar, S., Bjornsson, A., & Isaksson, L. A. (1994). The second to last amino acid in the nascent peptide as a codon context determinant. *EMBO J.* **13**, 249-257.

Nierhaus, K. H., Beyer, D., Dabrowski, M., Schäfer, M. A., Spahn, C. M. T., Wadzack, J., Bittner, J.-U., Burkhardt, N., Diedrich, G., Jünemann, R., Kamp, D., Voss, H., & Stuhrmann, H. B. (1995). The elongating ribosome: Structural and functional aspects. *Biochem.Cell Biol.* **73**, 1011-1021.

Nierhaus, K. H., Wadzack, J., Burkhardt, N., Jünemann, R., Meerwinck, W., Willumeit, R., & Stuhrmann, H. B. (1998). Structure of the elongating ribosome: Arrangement of the two tRNAs before and after translocation. *Proc.Natl.Acad.Sci.USA* **95**, 945-950.

Zar, J. H. (1996). The binomial test. *In* Biostatistical analysis, pp. 530-535. Upper Saddle River, New Jersey: Simon & Schuster.

CHAPTER VI

Identification and characterisation of the *Bacillus subtilis gtaC* gene, encoding the phosphoglucomutase involved in glucosylation of teichoic acid and phage susceptibility

VI.1. Summary

In the framework of the European *Bacillus subtilis* genome sequencing project an ORF, *yhxB*, was identified that displayed high homology to phosphoglucomutases and phosphomannomutases from both bacteria and eukaryotes. Phosphoglucomutase converts glucose-6-phosphate to glucose-1-phosphate. To study the possible function of the *yhxB* gene product, a Campbell-type mutant of *yhxB* with the concomitant formation of a promoter P_{yhxB}~*lacZ* fusion was constructed. The corresponding strain was subsequently analysed for growth, expression, phage ϕ25 and ϕ29 susceptibility and cell wall glucose content. We conclude from these analyses that *yhxB* encodes a phosphoglucomutase, probably involved in the glucosylation of teichoic acid, and that this gene corresponds to the *gtaC* marker. The *gtaC* (glucosylation of teichoic acid) marker was previously supposed to be located at around 77° on the *B. subtilis* 168 chromosome and to be responsible for the glucosylation of teichoic acid. We also conclude that the *B. subtilis* 168 chromosome does not encode a functional paralogue of *yhxB*. Gene *yhxB* is important for growth in glucose-based minimal medium, but not in nutrient broth. It is not essential for viability under the conditions tested.

VI.2. Introduction

Teichoic acids belong to the anionic polymers, a group of negatively charged polymers in the cell wall. Anionic polymers are divided into two classes: teichoic acids, in which a negative charge is provided by phosphodiester groups in the repeating units, and teichuronic acids, in which the negative charge is provided by the carboxyl groups of uronic acid residues.

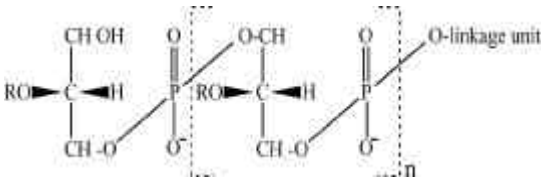


Fig.VI.1. Teichoic acid structure in *B. subtilis*. R = H, glucosyl or alanyl residues. From Archibald, 1993.

Teichoic acids vary in chemical composition between bacterial species and even between strains. In *Bacillus subtilis* 168, teichoic acid is a poly-1-*sn*-glycerol-3-phosphate in which the hydroxyl group at C-2 of the glycerol moiety may bear an α-glucosyl or D-alanyl ester substituent (Fig.VI.1.). The terminal phosphate bears a poly-

N-acetylhexosamine, linking it to an *N*-acetylmuramyl residue of the peptidoglycan (Archibald *et al.*, 1993). Besides peptidoglycan, the anionic polymers constitute a substantial proportion of the weight of walls of many gram-positive bacteria (Archibald *et al.*, 1993). Although it is probably essential for viability (Mauël *et al.*, 1989), the precise function of teichoic acid in the cell wall of *Bacillus* is as yet not entirely understood, and neither is the function of glucosylation of this compound. Available experimental data suggest that cells require anionic groups, rather than teichoic acid *per se*, for normal cell division and the development of the typical morphology of *B. subtilis* cells and colonies (Archibald *et al.*, 1993). From the work of Young (1967) it was known that *B. subtilis* 168 possesses a gene, *gtaC*, encoding the enzyme phosphoglucomutase, which is involved in the glucosylation of wall teichoic acid. Pooley and co-workers (1987) have demonstrated that *gtaC* is located close to *argC* on the genetic map of the *B. subtilis* chromosome, at about 77°C (Anagnostopoulos *et al.*, 1993). The *gtaC* gene is possibly accompanied by an additional, regulatory gene, *gtaE*, controlling the levels of phosphoglucomutase (PGM) and UDPglucose pyrophosphorylase (Pooley *et al.*, 1987). This finding was based on differences between mutants from the same linkage group with respect to their phage resistance pattern and cell wall galactosamine content. In this paper, we describe the function of gene *yhxB* in *B. subtilis* with respect to susceptibility to phages ϕ25, ϕ29, and SP10, cell-wall glucose content, and growth in glucose-based minimal medium. We postulate that *yhxB* probably corresponds to the *gtaC* marker, and that the putative regulatory gene, *gtaE*, is not encoded by the genes that are in the vicinity of *yhxB*.

VI.3. Methods

Cloning and sequencing

The cloning and sequencing of *yhxB* has been described in a previous paper (Noback *et al.*, 1998).

Homology analysis and sequence alignments

Homology comparisons were carried out using the FASTA program (Pearson & Lipman, 1988), and multiple sequence alignments with program ClustalW at the EBI services homepage at http://www2.ebi.ac.uk/services.html (Higgins *et al.*, 1994).

Media and growth conditions

Strains were cultured in rich medium (TY), nutrient broth (NB) or minimal medium (MM). TY consists of 10 g/l tryptone, 5 g/l yeast extract, 5 g/l NaCl, and 0.1 mM $MnCl_2$ at pH 7.2. Nutrient Broth contains 8 g/l Difco Bacto nutrient broth, 0.25 g/l $MgSO_4.7H_2O$, 1 g/l KCl, 0.01 mM $MnCl_2$, 0.001 mM $FeSO_4$, and 10 mM $CaCl_2$, at pH 7.1.

Minimal Medium consists of Spizizen's minimal salts (Spizizen, 1958), supplemented with glucose (0.5%), casein hydrolysate (0.02%; Difco Laboratories, Detroit, USA), and L-tryptophane (20 µg/ml).

Transformation and competence

B. subtilis cells were made competent essentially as described by Bron and Venema (1972). *E. coli* cells were made competent and transformed by the method of Mandel and Higa (1970).

Isolation of DNA

B. subtilis chromosomal DNA was purified as described by Bron (1990). Plasmid DNA was isolated by the alkaline-lysis method of Ish-Horowicz and Burke (1981).

PCR

PCR reactions were performed using Expand polymerase (Boehringer, GmbH, Mannheim, Germany) using buffers supplied with the enzyme, and according to protocols supplied by the manufacturer.

Protein determination

Protein concentration was measured using the Bio-Rad protein assay (Bio-Rad Laboratories GmbH, München, Germany) and the protocols supplied by the manufacturer.

ß-Galactosidase assay

Culture samples of 1 ml were taken at appropriate time points, and the cells were harvested by centrifugation for 2 min at 12000 *g* and stored at -20°C until use. Cells were lysed by incubation for 20 min at 37°C in 500 µl Z buffer containing per liter 10.7 g $Na_2HPO_4 \cdot 2H_2O$ (0.06 M), 5.52 g $NaH_2PO_4 \cdot H_2O$ (0.04 M), 0.75 g KCl, 0.246 g $MgSO_4 \cdot 7H_2O$. Before use, DTT was added to a final concentration of 1 mM and 1/100 (*vol/vol*) of lysis solution (1 mg/ml DnaseI and 10 mg/ml lysozyme in water). The samples were then centrifuged for 2 min at 12,000 × g, and the supernatant was collected and stored on ice. Subsequently, 200 µl of the protein sample was mixed with 600 µl of Z-buffer, and 200 µl ONPG solution containing 0.1 M Na_2HPO_4, 0.1 M NaH_2PO_4 and 4 mg/ml ONPG was added. Samples were then incubated at 28°C until a yellow colour appeared. The reactions were stopped by adding 500 µl 1 M Na_2CO_3. Extinction of the samples was measured at 420 nm and activity in Miller Units (M.U.: nmol ONPG × min^{-1} × mg^{-1} protein) was calculated as follows:

(OD420 × 1.5)/(sample vol × T × 0.00486 × mg/ml protein).

Preparation of cell walls

A single colony was used to inoculate 1 litre of 2 × YT and the cells were grown overnight at 37°C with shaking. The cells were pelleted by centrifugation, and the supernatant discarded. The cells were then resuspended in 200 ml of 1M MES pH 6.5, 0.4% (*w/v*) SDS. The cell suspension was subsequently boiled for 15 min, and then washed three times in 500

ml of 1 M MES pH 6.5. The cell walls were then resuspended in 10 ml of water, freeze-dried overnight, and stored at 4°C.

Removal of glucose from cell wall preparations

A defined amount of cell wall (~20 mg) was resuspended in 1.5 ml of 2M HCl and boiled for three hours. The sample was diluted to 10 ml with water and then freeze-dried overnight. The hydrolysed walls were washed three times by resuspension in 5 ml of water and freeze-dried overnight. Finally, the samples were resuspended in 1 ml of water and stored at 4°C. As a control, 1 mg of glucose was treated as a regular sample to calculate the fraction of glucose lost during the hydrolysis procedure.

Assay for glucose content using the glucose oxidase reaction

Glucose oxidase reagent was prepared by resuspending 1 PGO tablet (peroxidase and glucose oxidase from Sigma-Aldrich Chemie, Zwijndrecht, the Netherlands) in 100 ml of water to which ABTS (diammonium 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonate), Boehringer) was added to a final concentration of 1 mg/ml. Reactions were prepared by mixing 100 µl of sample with 5 ml of glucose oxidase reagent, and the reaction was allowed to continue for 45 minutes at room temperature in the dark, after which the absorbance of the samples was determined at 450 nm. Water was taken as the blank, and 100 µg of glucose resuspended in 100 µl water as standard. The concentration of glucose in the sample was calculated as follows: (Absorbance of sample at 450 nm)/(Absorbance of standard at 450 nm) × (concentration of standard).

Phage titre determination

The phage titre was determined by mixing 0.1 ml of phage suspension with 0.2 ml of *B. subtilis* culture grown in nutrient broth and 2 ml of molten soft agar (1%) at 45ºC. The mixture was then vortexed for 10 seconds, poured onto a nutrient agar plate and the plate was incubated overnight at 37ºC.

VI.4. Results and discussion

Similarity analysis and regulon identification

The deduced protein sequence of the *yhxB* gene was compared to all sequences in public databases, as well as to all available *B. subtilis* protein sequences. Besides numerous orthologs, one putative *B. subtilis* paralog of YhxB was identified, YbbT. However, its sequence is not included in the multiple sequence alignment presented in Fig. VI.2 because, although its sequence is 25 % identical to *yhxB*, it lacks 24 of the 59 conserved amino acids of the phosphogluco- and phosphomannomutases. In this figure, for clarity reasons, only orthologs with the highest similarity are included, since there are at least forty sequences with significant homology to the YhxB protein sequence. The catalytic site residue, the serine at position 146 (in YhxB), forms the phosphoserine intermediate.

Fig.VI.2. Multiple sequence alignment of (putative) phosphomanno- and phosphoglucomutases. Completely conserved residues are indicated by an asterisk; highly conserved residues with a double dot, and weakly conserved ones with a single dot. Ml: *Mycobacterium leprae* putative phosphomanno-mutase (PMM); Mt: *Mycobacterium tuberculosis* putative PMM/PGM (phosphogluco-mutase); Bs: *Bacillus subtilis* YhxB; Bb: *Borrelia burgdorferi* PMM; MP: *Mycoplasma pirum* PMM; Mg: *Mycoplasma genitalium* PMM.

YhxB is probably a monocistronic gene, since it is preceded and followed by a putative rho-independent terminator. The calculated free energies of these terminators are (at 37°C) -15.1 kCal (upstream of *yhxB*), and -14.8 kCal (downstream of *yhxB*). The 5' upstream region contains a putative σ^A-dependent promoter: GTGACA- 15nt –TATAAA.

Mutant construction

A Campbell-type mutant with concomitant transcriptional fusion to a *lacZ* reporter gene was constructed using plasmid pMUTin2 (Vagner *et al.*, 1998), which is the standard vector used for mutant construction and analysis in the European *B. subtilis* functional analysis program (Fig.VI.3.) An internal fragment of the *yhxB* coding region was amplified by PCR

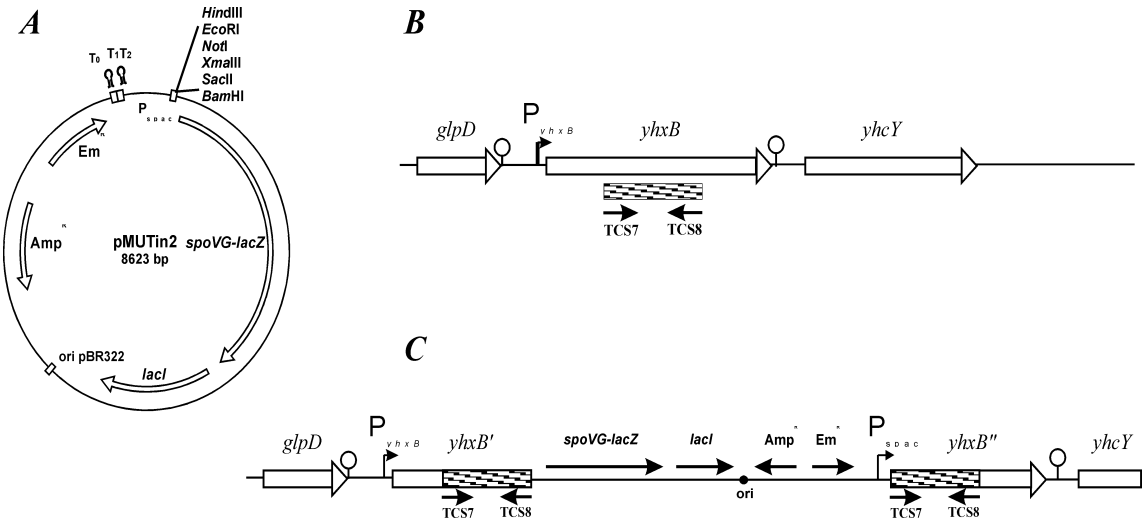


Fig.VI.3. Construction of a Campbell-type mutant yielding a *yhxB* knock-out and P_{yhxB}~*lacZ* fusion. Indicated are the map of the integrational plasmid pMUTin2 (A), the chromosomal organisation in the *yhxB* region with the fragment that was amplified by PCR and cloned into pMUTin2, and the resulting situation (strain TCS789) after integration of the pMUTin2 derivative in *yhxB* (C).

using primers TCS7 (*Hin*dIII tag; 5'-GCCGAAGCTTGTCAGACGCGGT-CTTGAA-3') and TCS8 (*Bam*HI tag; 5'-CGCGGATCCATTGATACGCTGACAGGCT-3'), and this fragment was cloned in pMUTin2. The resulting construct was used to transform *B. subtilis* 168 and transformants (Em^r colonies) were checked for correct integration of the pMUTin2 vector by PCR analysis and Southern hybridisation (data not shown). The resulting correct strain, *B. subtilis* TCS789, was used for all further analyses.

Growth, expression, and standard functional analysis

Since the *yhxB* mutant strain was viable, we concluded that gene *yhxB* is not essential. Growth and ß-galactosidase activity were determined for the wild-type strain, *B. subtilis* 168, and the *yhxB* mutant, *B. subtilis* TCS789. The assays were carried out in minimal medium and nutrient broth and the data are summarised in Fig.VI.4. The results show that growth was slightly impaired when the *yhxB* mutant was grown in minimal medium. In this medium, the

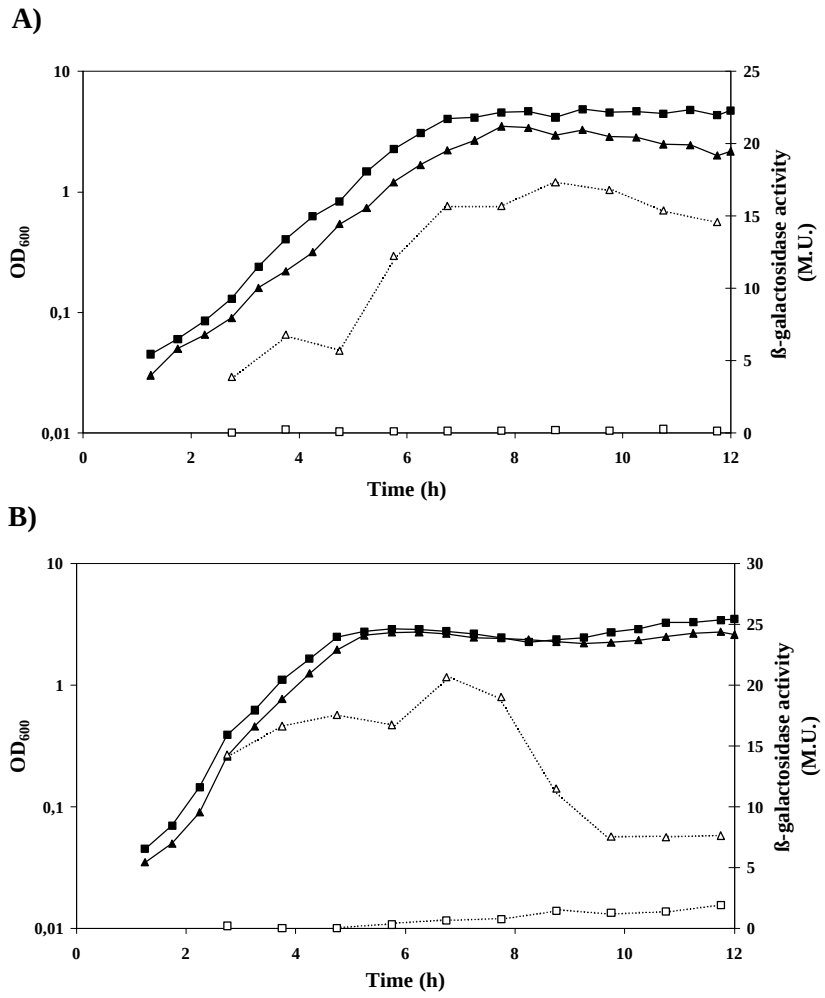


Fig. VI.4. Growth- and expression analysis of the *yhxB* mutant as compared to the wild-type. A): Minimal medium. B): Nutrient broth. Filled squares: growth of wild-type; filled triangles: growth of *yhxB* mutant; open squares: ß-galactosidase activity in wild-type; open triangles: ß-galactosidase activity of the *yhxB* mutant. Growth was expressed as optical density at 600 nm. ß-galactosidase activity was expressed in Miller Units (M.U.; nmol ONPG × min^{-1} × mg^{-1} protein).

mutant grew with a doubling time of 34.4 min as compared to 28.8 min for the wild-type, it reached a lower final cell density, and started to lyse already in early stationary phase. These differences were not observed in nutrient broth. Although *yhxB* was expressed during growth, its expression was highest around the switchpoint between logarithmic and stationary growth.

The *yhxB* mutant strain was also subjected to several tests that are employed in the European *Bacillus subtilis* functional analysis program, the aim of which is to analyse (and categorise) a large number of *B. subtilis* ORFs with unknown function with respect to processes such as protein secretion, sporulation, competence, recombination (measured as mitomycin C sensitivity), and stress responses. None of these tests revealed differences between the wild-type and *yhxB* mutant. Only during growth at increased (51°C) temperature, a difference was observed: although the *yhxB* mutant formed colonies of similar size as the wild-type strain, the centre of these colonies rapidly lysed after overnight incubation. No differences were observed during colony growth at 37°C or 15°C.

Phage susceptibility and cell wall glucose content

In order to demonstrate that *yhxB* corresponds to the *gtaC* marker, the *yhxB* mutant was analysed for cell wall glucose content and phage Φ25 and Φ29 susceptibility. In strain

TCS789, cell wall glucose content was reduced almost tenfold as compared to the wild-type strain, and the mutation in *yhxB* rendered the cells fully resistant to both phages, Φ25 and Φ29 (Table VI.1.).

Table VI.1. Cell wall glucose content and phage Φ25/Φ29 susceptibility

Strain	**cell wall glucose (μg/mg)**	**phage Φ25/Φ29 susceptibility** [$]
B. subtilis 168	163	normal; 10^9 p.f.u./ml
B. subtilis TCS789	18	zero

$: Determined by phage adsorption to the cells in liquid rich medium and subsequent plating on rich medium as an overlay in top agar. The titre after overnight incubation at 37°C was used as a measure for phage susceptibility.

In this paper, we have shown that a knock-out mutation could be constructed in ORF *yhxB*, indicating that this gene is not essential. ORF *yhxB* probably encodes the enzyme phosphoglucomutase. In the *yhxB* mutant strain, the cell wall glucose content and phage Φ25/Φ29 susceptibility were phenotypically similar to the corresponding properties of a phosphoglucomutase-deficient strain that was previously described by Young (1967). Although *gtaC* was previously mapped at 77°, and *yhxB* is located at 86° on the *B. subtilis* chromosome (Anagnostopoulos *et al.*, 1993), the difference between these positions is in all probability only seeming, as the physical/genetic map of this region of the *B. subtilis* chromosome has recently been thoroughly revised (see chapter II of this thesis). Therefore, we conclude that ORF *yhxB* most likely corresponds to the *gtaC* marker, responsible for glucosylation of teichoic acid.

We have also investigated the possibility that one of the genes in the vicinity of *yhxB* could encode the other teichoic acid marker, *gtaE*, the presence of which was postulated by Pooley and coworkers (1987). Upstream of *yhxB*, the *glpPFKD* operon is located, which is involved in uptake and catabolism of glycerol. These are not likely candidates, since *gtaE* was postulated to be a regulator of phosphoglucomutase and UDP-glucose pyrophosphorylase. Downstream of *yhxB*, we identified by similarity analysis more likely candidates: a putative two-component system of unknown function encoded by *yhcY* (sensory histidine kinase) and *yhcZ* (*degU*-like regulator). This two-component system is organised in an operon structure with a third ORF of unknown function which has one ortholog in *Escherichia coli*, *yieF*, also of unknown function. Insertional mutants of these three genes were constructed in essentially the same manner as was done for *yhxB*, and these were analysed for phage Φ25 and Φ29 susceptibility. The three mutant strains all displayed wild-type sensitivity to both bacteriophages. We conclude from these experiments that neither the two-component system encoded by *yhcYZ*, nor *yhdA* correspond to the *gtaE* marker.

References

Anagnostopoulos, C., Piggot, P. J., & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. *In Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, pp. 425-461. Edited by A. L. Sonenshein, J. A. Hoch, and R. Losick. Washington, DC: American Society for Microbiology.

Archibald, A. R., Hancock, J. M., & Harwood, C. R. (1993). Cell wall structure, synthesis, and turnover. *In Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, pp. 381-410. Edited by A. L. Sonenshein, J. A. Hoch, and R. Losick. Washington DC: American Society for Microbiology.

Bron, S. (1990). Plasmids. *In* Molecular Biological Methods for *Bacillus*, pp. 75-174. Edited by C. R. Harwood and S. M. Cutting. Chichester: John Wiley & Sons.

Bron, S. & Venema, G. (1972). Ultraviolet inactivation and excision repair in *Bacillus subtilis*. I. Construction and characterization of a eightfold auxotrophic strain and two ultraviolet-sensitive derivatives. *Mutat.Res.* **15**, 1-10.

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Ish-Horowicz, D. & Burke, F. J. (1981). Rapid and efficient cosmid cloning. *Nucleic Acids Res.* **9**, 2989-2999.

Mandel, M. & Higa, A. (1970). Calcium-dependent bacteriophage DNA infection. *J.Mol.Biol.* **53**, 159-162.

Mauël, C., Young, M., Margot, P., & Karamata, D. (1989). The essential nature of teichoic acids in *Bacillus subtilis* as revealed by insertional mutagenesis. *Mol.Gen.Genet.* **215**, 388-394.

Noback, M. A., Holsappel, S., Kiewiet, R., Terpstra, P., Wambutt, R., Wedler, H., Venema, G., & Bron, S. (1998). The 172 kb *prkA-addAB* region from 83° to 97° of the *Bacillus subtilis* chromosome contains several dysfunctional genes, the *glyB* marker, many genes encoding transporter proteins, and the ubiquitous *hit* gene. *Microbiol.* **144**, 859-875.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc.Natl.Acad.Sci.USA* **85**, 2444-2448.

Pooley, H. M., Paschoud, D., & Karamata, D. (1987). The *gtaB* marker in *Bacillus subtilis* 168 is associated with a deficiency in UDPglucose pyrophosphorylase. *J.Gen.Microbiol.* **133** , 3481-3493.

Spizizen, I. (1958). Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc.Natl.Acad.Sci.USA* **44**, 1072-1078.

Vagner, V., Dervyn, E., & Ehrlich, S. D. (1998). A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiology* **144**, 3097-3104.

Young, F. E. (1967). Requirement glucosylated teichoic acid for adsorption of phage in *Bacillus subtilis* 168. *Proc.Natl.Acad.Sci.USA* **58**, 2377-2384.

CHAPTER VII

The *Bacillus subtilis* counterpart of the ubiquitous Hit protein family is involved in heat-shock protection and hydrolyses ADP

VII.1. Summary

In order to determine the function of the *Bacillus subtilis hit* gene, two Campbell-type mutants were constructed. In one of these, the *hit* gene was inactivated and, concomitantly, the *lacZ* gene was transcriptionally fused to the *hit* promoter. In the other mutant, the *hit* gene was placed under control of the IPTG-inducible P_{spac} promoter with the *hit* promoter fused to *lacZ*. These mutants were used to assess a possible role of the *hit* gene in cellular processes such as growth, cell division, sporulation, protein secretion, competence, heat- and cold-resistance, and recombination. In the *B. subtilis* wild-type, transcription of the *hit* gene was high during logarithmic growth, but almost completely absent during stationary growth. No clear phenotype was observed with respect to any of the functions tested, with the exception of a moderate effect on heat-sensitivity. The Hit protein was purified by the histidine-tagging strategy, and biochemically characterised *in vitro*. The enzyme appeared to have ADP-hydrolysing activity *in vitro*. This contrasts with other members of the Hit family of proteins, several of which have been shown to be AppppA or ApppA hydrolases. The significance of this activity of *B. subtilis* Hit for the cell is unclear. To examine whether transcription of the *hit* gene is subject to regulation, various possible *B. subtilis* regulatory genes were inactivated, and their effect on *hit* transcription was examined. Northern analysis suggested that *yabJ*, for which no function was known, probably is a negative regulator of the *hit* gene acting in the stationary growth phase.

VII.2. Introduction

Previously, McDonald & Walsh (McDonald & Walsh, 1985) described the purification of a 17 kDa Ca^{2+}-binding protein from bovine brain with an *in vitro* protein kinase C (PKC) inhibitory activity. Later, it was reported that this was not a high-affinity Ca^{2+}-binding protein, and neither was Ca^{2+} required for its protein kinase C inhibitory activity. By means of Western immunoblotting, the protein was found to be present in several bovine, murine, avian and human tissues (McDonald *et al.*, 1987). The protein was classified as a novel, zinc-binding protein without having the typical zinc finger signature of other zinc binding proteins (Pearson *et al.*, 1990). In this new family of enzymes, zinc appeared to be bound to a site

consisting of a typical triad of closely positioned histidine residues: His-X-His-X-His (Mozier *et al.*, 1991). The typical HIT motif was later refined to His-ϕ-His-ϕ-His-ϕ-ϕ, where ϕ is any hydrophobic amino acid (Brenner *et al.*, 1997). More proteins were subsequently found with this HIstidine Triad motif and the name Hit protein was introduced (Séraphin, 1992). In following years, many genes encoding Hit-like proteins, or PKCI-1 proteins (protein kinase C interacting proteins), were identified in organisms from archaeal, bacterial, lower- and higher eukaryotic origin. Among these are: *Synechococcus* sp. (Bustos *et al.*, 1990), maize (Simpson *et al.*, 1994), yeast (Fröhlich *et al.*, 1991), *B. subtilis* (Noback *et al.*, 1998), *Lupinus angustifolius* (Maksel *et al.*, 1998), and even the bacterium with the smallest genome known to date, *Mycoplasma genitalium* (Fraser *et al.*, 1995). In 1996, a human member of the *hit* gene family, *fhit* (**f**ragile **hi**stidine **t**riad), was identified and localised at the tumor-associated chromosomal fragile site FRA3B at 3p14.2, where it spans a DNA region of approximately 1 Mb. It was found to be aberrantly transcribed in a significant portion of digestive tract cancers (Ohta *et al.*, 1996), small- and non-small cell lung cancers (Sozzi *et al.*, 1996a). Many reports associating aberrant *fhit* transcripts with various types of (epithelial) carcinomas have since then been published: Merkel cell carcinomas (Sozzi *et al.*, 1996b), breast carcinomas (Negrini *et al.*, 1996; Panagopoulos *et al.*, 1996), non-comedo ductal carcinomas (Man *et al.*, 1996), colorectal carcinomas (Thiagalingam *et al.*, 1996), pancreatic tumors (Shridhar *et al.*, 1996), and lung cancer cell lines (Yanagisawa *et al.*, 1996). Human *fhit* is now widely recognised as a putative tumor-suppressor gene.

The Hit superfamily of proteins is now subdivided into two branches. One branch comprises the group of PKCI homologs that seems to be ubiquitous in nature. The *B. subtilis* Hit protein is a member of this group. The PKCI group is also referred to in the literature with the name HINT, which stands for histidine triad nucleotide-binding motif. The other branch, evolutionary divergent from the PKCI-1 branch, comprises the group of FHIT homologs which probably contains only eukaryotic members. The amino-terminal amino acid sequence of members of this group is, in contrast with the PKCI group of proteins, not much conserved. Humans possess representatives of both branches of the Hit superfamily of proteins: FHIT and PKCI-1. Mammalian members of the PKCI branch are characterised by a highly conserved C-terminal sequence: GGRXXXWPPG (Lima *et al.*, 1997; see also Fig. VII.1). Human PKCI-1 has been solved to high resolution by X-ray crystallography. In its active form it is a homodimer, with specific interaction of the conserved histidines with zinc (Gilmour *et al.*, 1997; Lima *et al.*, 1996).

The biochemical activity of some members of the Hit superfamily of proteins has been elucidated. Human FHIT protein has been identified *in vitro* as a dinucleoside $5',5'''$-P^1,P^3-triphosphate (Ap_3A) hydrolase generating ADP and AMP as reaction products (Barnes *et al.*, 1996). The *Schizosaccharomyces pombe* (gene *aph1*) and *Lupinus angustifolius* members of the Hit protein family were shown to be $5',5'''$-P^1,P^4-tetraphosphate (Ap_4A) asymmetrical hydrolases (Huang *et al.*, 1995; Maksel *et al.*, 1998). The human PKCI homolog has been shown to bind and hydrolyse ADP. Structure-based analysis of catalysis of human FHIT and PKCI Hit proteins has unified the Hit family as nucleotidyl hydrolases, transferases, or both. Substrate specificity is probably dictated by the specific composition of the C-terminal amino

acid residues (Lima *et al.*, 1997). However, the biological function(s) of this family of proteins still remains to be elucidated.

This paper deals with the search for a possible biological function of the *B. subtilis hit* gene. This was done through systematic phenotype screening, expression analysis via fusion to a reporter gene and by Northern blotting, and a search for possible regulators of the *B. subtilis hit* gene.

VII.3. Materials and methods

Strains and plasmids

The bacterial strains and plasmids used in this study are listed in Table VII.1.

Table VII.1. Plasmids and strains

Plasmid	Genotype	Comment/Reference
pMUTin2	*Em*^{R} *Ap*^{R}	Used for mutant construction (Vagner *et al.*, 1998)
pMUTH75	*Em*^{R} *Ap*^{R}	pMUTin2 carrying an internal fragment of *hit*; this chapter
pMUTH30	*Em*^{R} *Ap*^{R}	pMUTin2 carrying an N-terminal fragment of *hit*; this chapter
pSU2718	*Cm*^{R}	Martinez *et al.*, 1988
pDG792	*Km*^{R}	Guérout-Fleury *et al.*, 1995
pSK1	*Km*^{R} *Cm*^{R}	pSU2718 with *lacZ**a*** replaced by Km^R from pDG792; this chapter
pSKYJ1	*Km*^{R} *Cm*^{R}	pSK1 carrying an internal fragment of *yabJ*; this chapter
pMPH51	*Em*^{R} *Ap*^{R} P_{hit}*~lacZ*	this chapter

Strain	Genotype	Comment/Reference
B. subtilis 168	*trpC2*	wild-type strain
B. subtilis HVH75	*trpC2 hit* P*hit~lacZ Em*^{R}	pMUTin2 *hit* Campbell insertion mutant and *lacZ* fusion; this chapter
B. subtilis HVH30	*trpC2* P*hit~lacZ Em*^{R}	pMUTin2 *hit* Campbell with *lacZ* fusion; this chapter
B. subtilis HVH77	*trpC2 hit rpoE::CAT* P*hit~lacZ Cm*^{R} *Em*^{R}	HVH75 derivative; this chapter
B. subtilis HVH78	*trpC2 hit yabJ* P*hit~lacZ Km*^{R} *Em*^{R}	HVH75 derivative; *yabJ* Campbell insertion mutant with pSKYJ1; this chapter
B. subtilis YABJdd	*trpC2* Δ*yabJ* P*yabJ~lacZ Em*^{R}	pMUTin2-mediated deletion of *yabJ* by double crossing over; pers. comm. Dr. Kobayashi
E. coli BL21	(F$^-$ *ompT* r^-_Bm^-_B)	(Studier *et al.*, 1990)
E. coli BL21_HC1	*E. coli* BL21 carrying plasmid pT7HC1 (Hit~His$_6$) *Ap*^{R}	this chapter
E. coli XL1-Blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F' *proAB lacI*^{q}*Z M15* Tn*10* (*Tet*^{R})]	Stratagene (USA)
E. coli MHF1	*E. coli* XL1-Blue carrying plasmid pMHF1 (MBP~Hit fusion) *Ap*^{R} *Tet*^{R}	this chapter
E. coli MPH51	*E. coli* XL1-Blue carrying plasmid pMPH51 *Ap*^{R} *Em*^{R}	this chapter

Homology analysis and sequence alignments

Homology comparisons were carried out using the FASTA program (Pearson & Lipman, 1988), and multiple sequence alignments with the program ClustalW at the EBI services homepage at http://www2.ebi.ac.uk/services.html (Higgins *et al.*, 1994).

Media and growth conditions

Strains were cultured in trypton-yeast medium (TY), Difco sporulation medium (DSM), or minimal medium (MM), at 37°C unless stated otherwise. TY consists of 10 g/l tryptone, 5 g/l yeast extract, 5 g/l NaCl, and 0.1 mM $MnCl_2$ at pH 7.2. DSM contains 8 g/l Difco Bacto nutrient broth, 0.25 g/l $MgSO_4.7H_2O$, 1 g/l KCl, 0.01 mM $MnCl_2$, 0.001 mM $FeSO_4$, and 10 mM $CaCl_2$, at pH 7.1. Minimal medium consists of Spizizen's minimal salts (Spizizen, 1958), supplemented with glucose (0.5%), casein hydrolysate (0.02%; Difco Laboratories, Detroit, USA), and L-tryptophane (20 µg/ml).

Transformation and competence

B. subtilis cells were made competent essentially as described by Bron and Venema (1972). *E. coli* cells were made competent and transformed by the method of Mandel and Higa (1970). For transformation of plasmid genome shotgun banks to *E. coli*, the method of Inoue *et al.* (1990) was used.

Isolation of DNA

B. subtilis chromosomal DNA was purified as described by Bron (1990). Plasmid DNA was isolated by the alkaline-lysis method of Ish-Horowicz and Burke (1981).

PCR

PCR reactions were performed using Expand polymerase (Boehringer, GmbH, Mannheim, Germany) in buffers supplied with the enzyme, and according to protocols supplied by the manufacturer.

Protein determination

Protein concentrations were measured using the Bio-Rad protein assay (Bio-Rad Laboratories GmbH, München, Germany) and the protocols supplied by the manufacturer.

ß-Galactosidase assays

Culture samples of 1 ml were taken at appropriate time points, and the cells were harvested by centrifugation for 2 min at 12,000 × *g* and stored at -20°C until use. ß-galactosidase assays were carried out essentially as described by Miller (1982) and activity was expressed in Miller Units (M.U.: nmol ONPG × min^{-1} × mg^{-1} protein), calculated as follows:

(OD420 × 1.5)/(sample vol × T × 0.00486 × mg/ml protein).

Northern analysis

The Digoxigenin(DIG)-labelling method was used for Northern analysis, with reagents and protocols supplied by Boehringer Mannheim GmbH for labelling, hybridization, and detection. The 252 bp *hit* probe fragment was generated by PCR using primers hit-F (5'-ATGCATTGTGCAGAGAATTG-3') and hit-T7 (T7 promoter sequence tag 5'-TAATACGACTCACTATAGGGCGAGTGGAACACAGATTGTCCAG-3'). This PCR fragment was used as a template for the synthesis of DIG-labelled RNA probe by T7 RNA polymerase.

Protein purification

The Hit~His_6 fusion protein was purified using a Talon column, according to the protocols supplied by the manufacturer (Clontech Laboratories Inc., Palo Alto, Ca, USA).
The Hit~MBP (Maltose Binding Protein) fusion protein was purified with amylose resin using the Protein Fusion and Purification System (New England Biolabs GmbH, Schwalbach, Germany), as described by Guan *et al* (1987).

Nucleotide hydrolysis assay

With minor modifications, the method of Barnes *et al.* (1996) was used for determination of Hit activity. HPLC analysis was used to determine the substrate specificity and reaction products of the Hit~His_6 protein. HPLC and associated Model 1706 UV/Vis detection system were obtained from Bio-Rad (Bio-Rad Laboratories, Hercules, Ca, USA). Potential nucleotide substrates were incubated at concentrations of 80 μM with or without protein sample in 50 mM MES-Tris buffer (between pH 4 to pH 9), 0.5 mM $MnCl_2$ at 37°C for 30 min in a final reaction volume of 200 μl. Subsequently, 100 μl of the reaction mixture was injected onto a MonoQ HPLC column. The products were then eluted using a gradient of 25 to 600 mM NH_4HC0_3, pH 8.5. Nucleotides were detected at 254 nm and identified by retention time. Peak areas were integrated using the 'valuechom' software from Bio-Rad. The effect of different divalent cations was determined by substituting $MnCl_2$ in the reaction mixture for 0.5 mM of the cation to be tested.

Construction of mutants

Two Campbell-type mutants of the *hit* gene were constructed using plasmid pMUTin2, which is the standard vector used for mutant construction and analysis in the European *B. subtilis* gene function analysis program (Vagner *et al.*, 1998). An insertion mutant with concomitant transcriptional fusion of the *hit* promoter to the *lacZ* reporter gene was constructed using an internal fragment of the *hit* coding region (see also Fig. VI.3). The internal fragment was generated using primers HN07 (*Hin*dIII tag; 5'-GCCGAAGCTTAGC-CAAGTGACAAAAG-G-3') and HC02 (*Bam*HI tag; 5'-CGCGGATCCAACACAGATTGT-CCAGC-3'). Using the same pMUTin2 plasmid, another integrant was constructed with a transcriptional fusion of the *hit* promoter to the *lacZ* reporter gene and the intact *hit* gene under control of the *spac* promoter (thus IPTG inducible). This was done using an N-terminal fragment of the *hit* coding region which was amplified with primers HN07 (see above) and

HRFB (*Bam*HI tag; 5'-CGCGGATCCTTATGATGAGGCCAGGC-3'). Both fragments were cloned in pMUTin2, and the resulting construct was used to transform *B. subtilis* 168 and transformants (Em^r colonies) were checked for correct integration of the pMUTin2 vector by PCR analysis and Southern hybridisation (data not shown). The resulting correct strains, *B. subtilis* HVH75 (insertion mutant; *lacZ* fusion) and HVH30 (*lacZ* fusion; IPTG inducible), were used for further analyses.

A mutant of the *yabJ* gene was constructed as follows. First, a plasmid was constructed without pMUTin2 sequences. This was achieved by replacing the *lacZ**a*** containing *Hae*II restriction fragment of plasmid pSU2718 (Martinez *et al.*, 1988) with the *Bam*HI-*Stu*I Km^r containing restriction fragment from plasmid pDG792 (Guérout-Fleury *et al.*, 1995). The resulting plasmid was named pSK1 (not shown). An internal fragment of *yabJ* was amplified by PCR using primers PyabjF (*Pst*I tag; 5'-AAAACTGCAGAAATGGTGAATGGCGATA-3') and PyabjR (*Eco*RI tag; CCGGAATTCTTCCGCAAACTGTTCCATA-3'). This fragment was cloned in plasmid pSK1, and the resulting correct plasmid pSKYJ1 was used to transform *B. subtilis* HVH75. Kanamycin/erythromycin resistant transformants were verified by PCR and Southern hybridization, and the resulting correct *yabJ/hit* double-mutant strain was named HVH78.

In order to purify and biochemically characterise the Hit protein, an *Escherichia coli* mutant strain was constructed carrying a *hit* gene encoding a carboxy-terminal histidine-tagged variant of Hit. The mutant gene was obtained via PCR using primers HN02 (*Hin*dIII tag; 5'-GCCGAAGCTTAATTTCATAAGGAGGAACCCTCATG-3') and HHRC1 (*Bam*HI tag; **6×His** tag 5'-CGCGGATCCTTA**GTGATGGTGATGGTGATG**TGATGAGGCCAG-G-3'). After PCR, the DNA fragment with the mutant gene was cloned into the *E. coli* expression vector pT712. The construct was used to transform *E. coli* BL21. Plasmid was isolated from transformants (Ap^r colonies) and the integrity of the construct was verified by PCR and sequence analysis. The resulting *E. coli* strain with the correct plasmid, pT7HC2, was used to purify his-tagged Hit protein.

VII.4. Results and discussion

Sequencing and homology analysis

We have previously reported the cloning and sequencing of the *B. subtilis hit* gene (Noback *et al.*, 1998). This gene was formerly designated *yhaE*, following the standard nomenclature employed by participants of the *B. subtilis* sequencing project for newly identified genes with unknown function. An overview of the chromosomal region in which the *B. subtilis hit* gene resides is presented in Fig.VII.6B.

In Fig.VII.1A, a multiple sequence alignment of known members of the Hit protein family is presented. Fig.VII.1B shows the corresponding phylogenetic tree. The phylogenetic tree shows that the two human representatives of the Hit superfamily, PKCI-homolog and FHIT, are not grouped together. The PKCI-homolog is placed into a group with only eukaryotic members, while FHIT is placed into a group with mainly prokaryotic members. This is likely reflective of an early divergence of these human Hit proteins.

```
BosT     ----------------ADEIAKAQVARP------------GGDTIFGKIIRKEIPAKI 30
RatN     ----------------MADEIAKAQVARP------------GGDTIFGKIIRKEIPAKI 31
H_PKCI   ----------------XADEIAKAQVARP------------GGDTIFGKIIRKEIPAKI 31
MusM     ----------------MADEIAKAQVAQP------------GGDTIFGKIIRKEIPAKI 31
CeaE     ----------------MSEVDKAHLAAINKD--------VQANDTLFGKIIRKEIPAKI 35
SynP     ----------------------------------------MSEDTIFGKIIRREIPADI 19
SynS     ----------------------------------------MAEDTIFSKIIRREIPAAI 19
BraJ     ------------------------------------------DTIFGKIISKEIPSTV 16
ZeaM     ----------------MSSEKEAALRRL------------DDSPTIFDKIIKKEIPSTV 31
HaeI     ----------------------------------------MAEETIFSKIIRKEIPANI 19
MycG     ----------------------MEKN-------------TTSSCIFCDIVQGSITSYK 23
MycP     ---------------------MVQKQSM------------ANNNCIFCGIVEGNVKSFK 26
H_FHIT   ------------------------------------------MSFRFGQHLIKPSVV 15
SchP     ----------------------------------------MPKQLYFSKFPVG-SQV 16
AzoB     ------------------------MAKT------------YDPNNVFARILRGEIPCKK 23
BacS     ------------------------------M--------HCAENCIFCKIIAGDIPSAK 21
MetJ     ------------------------------------------MCIFCKIINGEIPAKV 16
MycH     ------------------------------M--------NNWQEELFLKIIKREEPATI 21
SaCC     ------------------------MSAP--A--------TLDAACIFCKIIKSEIPSFK 25
MycT     MSAPSASIRACAAAPAKARSVGSTAAAYPVHLGGLREVQYRSDMPCVFCAIIAGEAPAIR 60
MycL     ---------------------MHLALWGRATQTALRHAYAEAMATIFTKIINRELPGRF 38
                                                             :

BosT     IYEDDQCLAFHDISPQAPTHFLVIPKKY--ISQISAAEDDDESLLGHLMIVGKKCAADLG 88
RatN     IYEDDQCLAFHDISPQAPTHFLVIPKKY--ISQISAAEDDDESLLGHLMIVGKKCAADLG 89
H_PKCI   IFEDDRCLAFHDISPQAPTHFLVIPKKH--ISQISVAEDDDESLLGHLMIVGKKCAADLG 89
MusM     IFEDDRCLAFHDISPQAPTHFLVIPKKH--ISQISVADDDDESLLGHLMIVGKKCAADLG 89
CeaE     IFEDDEALAFHDVSPQAPIHFLVIPKRR--IDMLENAVDSDAALIGKLMVTASKVAKQLG 93
SynP     VYEDDLCLAFRDVAPQAPVHILVIPKQP--IANLLEATAEHQALLGHLLLTVKAIAAQEG 77
SynS     VYEDDLCLAFKDVNPQAPVHVLLIPKKP--LPQLSAATPEDHALLGHLLLKAKEVAADLG 77
BraJ     VYEDDKVLAFRDITPQGPVHILLIPKVRDGLTGLFKAEERHIDILGRLLYTAKLVAKQEG 76
ZeaM     VYEDEKVLAFRDINPQAPTHILIIPKVKDGLTGLAKAEERHIEILGYLLYVAKVVAKQEG 91
HaeI     VYQDELVTAFRDISPQAKTHILIIPNKV--IPTVNDVTEQDEVALGRLFSVAAKLAKEEG 77
MycG     IGENEHAIAFLDAFPVADGHTLVIPKKH---AVDFSSTDQKELQAVSLLAKQIALKLKMT 80
MycP     VGENEHAFAFLDAFPVADGHTLVIPKKH---AVNYSSTDDESLKAVSLLAKEMALKLQQR 83
H_FHIT   FLKTELSFALVNRKPVVPGHVLVCPLRP--VERFHDLRPDEVADLFQTTQRVGTVVEKHF 73
SchP     FYRTKLSAAFVNLKPILPGHVLVIPQRA--VPRLKDLTPSELTDLFTSVRKVQQVIEKVF 74
AzoB     VLETEHALAFHDINPQAPTHILVIPKGAY-VDMDDFSARATEAEIAGLFRAVGEVARGAG 82
BacS     VYEDEHVLAFLDISQVTKGHTLVIPKTH--IENVYEFTDELAKQYFHAVPKIARAIRDE- 78
MetJ     VYEDEHVLAFLDINPRNKGHTLVVPKKH--YERFDEMPDDELCNFIKGVKKTVEVLKKLG 74
MycH     LYEDDKVIAFLDKYAHTKGHFLVVPKNYS-RNLFSISDEDLSYLIVKAREFALQEIKKLG 80
SaCC     LIETKYSYAFLDIQPTAEGHALIIPKYHG--AKLHDIPDE---FLTDAMPIAKRLAKAMK 80
MycT     IYEDGGYLAILDIRPFTRGHTLVLPKRH--TVDLTDTPPEALADMVAIGQRIARAARATK 118
MycL     VYEDDDVVAFLTIEPMTQGHTLVVPCAE-----IDQWQNVDPAIFGRVIAVSQLIGKGVC 93
         . .     *:          * *: *

BosT     LKK-GYRMVVNEGSDGGQSVYHVHLHVLGGR--------QMNWPPG-------------- 125
RatN     LKK-GYRMVVNEGSDGGQSVYHVHLHVLGGR--------QMNWPPG-------------- 126
H_PKCI   LNK-GYRMVVNEGSDGGQSVYHVHLHVLGGR--------QMHWPPG-------------- 126
MusM     LKR-GYRMVVNEGADGGQSVYHIHLHVLGGR--------QMNWPPG-------------- 126
CeaE     MAN-GYRVVVNNGKDGAQSVFHLHLHVLGGR--------QLQWPPG-------------- 130
SynP     LTE-GYRTVINTGPAGGQTVYHLHIHLLGGR--------SLAWPPG-------------- 114
SynS     IGD-QFRLVINNGAEVGQTVFHLHLHILGGR--------PFSWPPG-------------- 114
BraJ     LDE-GFRIVINDGPQGCQSVYHIHVHLIGGR--------QMNWPPG-------------- 113
ZeaM     LED-GYRVVINDGPSGCQSVYHIHVHLLGGR--------QMNWPPG-------------- 128
HaeI     VAEDGYRLIVNCNKHGGQEVFHLHMHLVGGE--------PLGRMLAK-------------- 116
MycG     LKPSGLNYVSNEGAIAGQVVFHFHLHIVPKY--------ETGKGFGYNVNKTNKRSL---- 129
MycP     LQPAGLNYVVNEGAKAGQEVFHYHMHVVPKY--------ETGLGFCYNVRKTNNRSI---- 132
H_FHIT   HGT-SLTFSMQDGPEAGQTVKHVHVHVLPRK--------AGDFHRNDSIYEELQKHD---- 121
SchP     SAS-ASNIGIQDGVDAGQTVPHVHVHIIPRK--------KADFSENDLVYSELEKNEGNL 125
AzoB     AAEPGYRILSNCGEDANQEVPHLHIHVFAGR--------RLGPMITKG------------ 122
BacS     FEPIGLNTLNNNGEKAGQSVFHYHMHIIPRY--------GKGDGFGAVWKTHADDYK---- 127
MetJ     FD--GYNIVNNNGRVAGQEVNHVHFHIIPRYXGD-GEVVKFGEVKNVDLDEVLKEIKG-- 129
MycH     AT--GFKLLINNEPDAEQSIFHTHVHIIPYYKK-------------------------- 111
SaCC     LDT--YNVLQNNGKIAHQEVDHVHFHLIPKRDEK-SG-LIVGWPAQETDFDKLGKLHKEL 136
MycT     LAD-ATHIAINDGRAAFQTVFHVHLHVLPPRNGD-----KLSVAKGMMLRRDPDREATGR 172
MycL     RA---FNAERAGVIIAGFEVPHLHIHVFPTHSLSNFSFANVDRNPSPESLDAAQDKIKAA 150
                         : * *.*:.

MycG     ----------------EENYQ--------------LISESKN---------------- 141
MycP     ----------------EANWE--------------LLTKEVD---------------- 144
H_FHIT   -------------KEDFPASWR-----------SEEEMAAEAAALRVYFQ--------- 147
SchP     ASLYLTGNERYAGDERPPTSMRQAIPKDEDRKPRTLEEMEKEAQWLKGYFSEEQEKE--- 182
BacS     ----------------PEDLQN----------ISSSIAKRLASS--------------- 145
SaCC     LAKLEGSD---------------------------------------------------- 144
MycT     ILREALAQQDAAAQD--------------------------------------------- 187
MycL     LTQLA------------------------------------------------------- 155
```

B

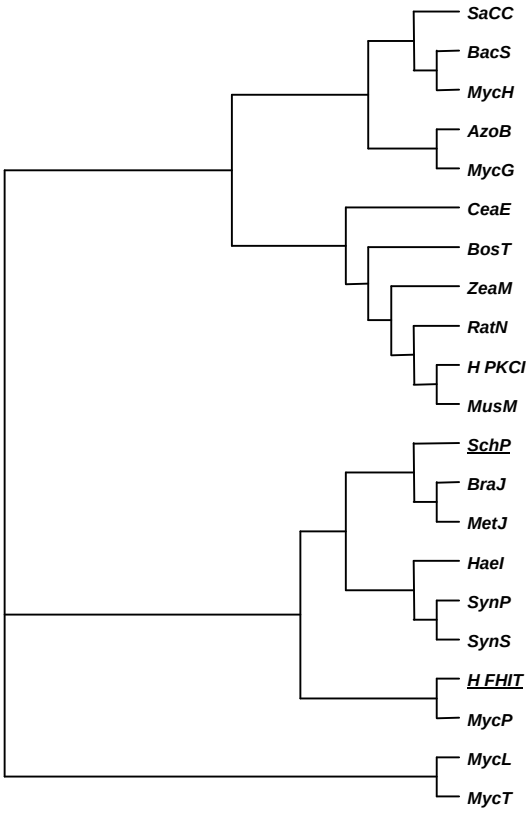


Fig.VII.1. Sequence alignment (**A**; previous page) and dendrogram (**B**) of members of the HIT protein family. In (**A**), completely conserved residues are indicated by an asterisk, highly conserved residues with a double dot, and weakly conserved ones with a single dot. In (**B**), HIT proteins that are not of the PKCI family are underlined. BosT: *Bos taurus* PKCI; RatN: *Rattus norvegicus* PKCI; H FHIT: *Homo sapiens* FHIT; H PKCI: *Homo sapiens* PKCI; MusM: *Mus musculus* PKCI; CeaE: *Caenorhabditis elegans* PKCI; SynP: *Synechococcus sp.* PKCI; SynS: *Synechocystis sp.* PKCI; BraJ: *Brassica juncea* PKCI; ZeaM: *Zea mays* PKCI; HaeI: *Haemophilus influenzae* PKCI; MycG: *Mycoplasma genitalium* PKCI; MycP: *Mycoplasma pneumoniae* PKCI; SchP: *Schizosaccharomyces pombe* Ap_4A hydrolase (*aph1*); AzoB: *Azospirillum brazilenze* PKCI; BacS: *Bacillus subtilis* PKCI; MetJ: *Methanococcus jannaschii* PKCI; MycH: *Mycoplasma hyorhinis* PKCI; SacC: *Saccharomyces cerevisiae* PKCI; MycT: *Mycobacterium tuberculosis* PKCI; MycL: *Mycobacterium leprae* PKCI.

Growth, expression and functional analysis tests

In order to assign a possible phenotypically recognisable function to the *B. subtilis hit* gene, growth and expression characteristics were determined for the mutant strains HVH75 and HVH30 as compared to the wild-type strain *B. subtilis* 168. These data are presented in Fig.VII.2 for growth in sporulation medium (DSM). Growth and expression were also determined in minimal and trypton-yeast extract media (MM and TY; data not shown). In these media growth was, as in DSM, unaffected in the mutants. Expression of *hit* seems to be growth-phase dependent; it is highest around the switch-point between logarithmic growth-

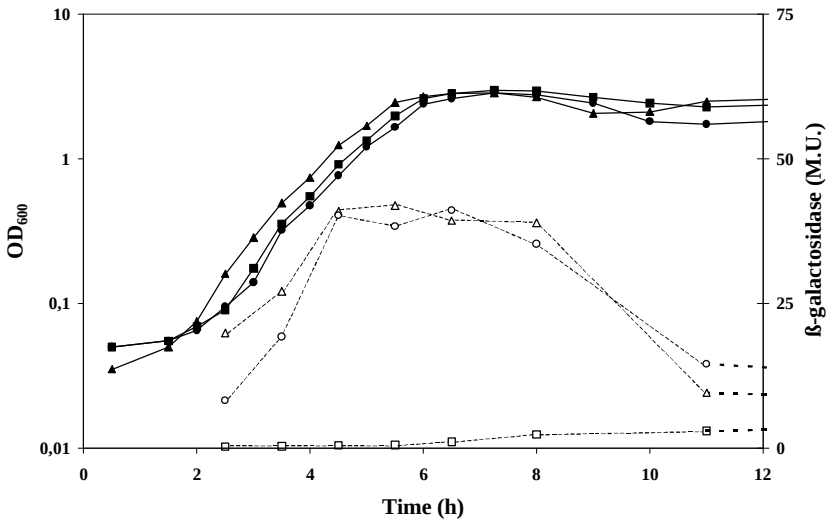


Fig.VII.2. Growth- and expression analysis of *hit* mutants in nutrient broth. Filled squares: growth of wild-type strain *B. subtilis* 168; filled triangles: growth of HVH75; filled circles: growth of HVH30. Open symbols represent the ß-galactosidase activity (in Miller Units) of the corresponding strain.

and the stationary growth phase. The expression patterns were also similar in DSM and MM, but the level of expression was higher during growth in TY and MM, with a maximum of about 110 M.U. Since expression levels were similar in both strains, HVH75 and HVH30 (grown in the presence of IPTG), we concluded that *hit* gene transcription is not subject to autoregulation.

We have subjected the *hit* mutants to the standard functional analysis tests that are employed by our group in the European *B. subtilis* functional analysis program. The *hit* mutants were screened for the following possible phenotypes: sporulation, germination, production and secretion of α-amylase and proteases, levansucrase activity, growth in reducing conditions (β-mercaptoethanol), anaerobic growth, growth at elevated and lowered temperatures (15°C and 48°C), recombination (sensitivity to mitomycin C), growth with iron depletion (in the presence of EDDA), survival after freezing, survival after u.v. irradiation, and survival after heat-shock treatment. In these assays, only heat-shock treatment showed a moderate difference between the mutant HVH75 and *B. subtilis* 168. We determined the percentage of cells, from logarithmically growing cultures of *B. subtilis* 168 and HVH75 (at 37°C), that survived after incubation for respectively one, five, and ten minutes at 50°C (see Fig.VII.3).

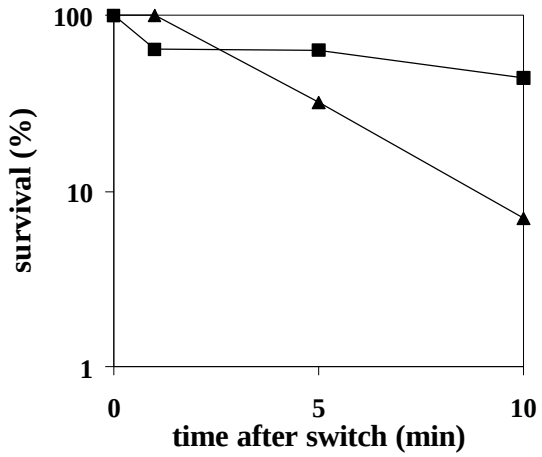


Fig.VII.3. Survival of HVH75 and *B. subtilis* 168 after heat-shock treatment (switch from 37°C to 50°C), as determined by viable count. Squares: *B. subtilis* 168; triangles: *B. subtilis* HVH75.

Identification of a negative regulator of *hit*; Northern analysis of *hit* transcripts in wild-type and *yabJ* background

With a shotgun cloning experiment in *E. coli*, we have attempted to identify possible regulators of the *hit* gene. In Figure VII.4, the plasmid with which the shotgun cloning experiment was performed, pMPH51, is depicted. The *hit* promoter region was cloned directly upstream of the *lacZ* reporter gene of plasmid pMUTin2 using primers HN05 (*Hin*dIII tag; 5'-GCCGAAGCTTGTATATCCG-CCGGTCAAGT-3') and HC02 (*Bam*HI tag; 5'CGCG-GATCCAACACAGATTGTCCAGC-3'). The resulting construct, with the 823 bp *hit* promoter fragment, yielded light-blue colonies after transformation to *E. coli*. Subsequently, *B. subtilis* chromosomal DNA was shotgun-cloned as fragments of about 1,5 - 6 kb into the unique *Sma*I restriction site of pMPH51 by the DNaseI method (described in chapter II.3), and *E. coli* was transformed with this shotgun bank by the Inoue (1990) method. *E. coli*

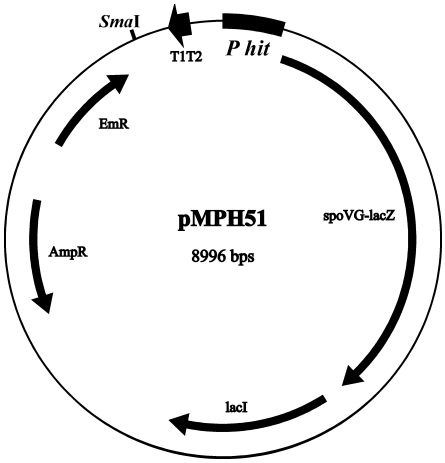


Fig.VII.4. Plasmid used for shotgun-cloning of regulators. P *hit*: *hit* promoter. T1T2: terminators. See text for details.

colonies showing altered blue colouring phenotypes were selected for further analysis. Since the sequence of the entire *B. subtilis* genome was available (Kunst *et al.*, 1997), it was feasible to sequence the insert endpoints of clones containing putative regulators and scan the genome for these sequences to determine the corresponding regions. Four clones possibly containing regulators of *hit* were identified in this manner. These are listed in Table VII.2.

Table VII.2. Clones containing possible regulators of *hit*

Clone	Colouring*	Co-ordinates of insert size†	Gene(s) on insert‡	Function
pMPH51	light blue			
pMPH51_A48		3904478..3908319	***sacT***	Positive regulator of *sacA* and *sacP*
	–	3841 bp		
			ywcI	Unknown
pMPH51_B6	± #	3810272..3812546	***ctrA***	CTP synthetase
		2274 bp	***rpoE***	RNA polymerase delta subunit
pMPH51_B47		53746..56833(nd)	***purR***	Purine operon repressor
	–	3087 bp	***yabJ***	Unknown; “ubiquitous”
			spoVG	Stage V sporulation
pMPH51_C36		3381388..3384224	***yusZ***	Unknown; similar to dehydrogenases
		2836 bp		
	+		***mrgA***	Metalloregulation; DNA binding protein
			yvtB	Unknown; similar to HtrA

***** clones that were lighter blue are indicated with –; clones that were darker blue with a +.
† Co-ordinates on the *B. subtilis* genome.
‡ Only intact genes are listed.
# This clone displayed a specific pattern of colouring: white in the centre of colonies and dark blue at the edges.

To verify whether the clones contained genes encoding regulators of the *hit* gene, mutants of two selected genes, *rpoE* and *yabJ*, were constructed in a *hit*⁻ mutant background. We selected these two genes for the following reasons. The *B. subtilis rpoE* gene has been shown to be able to reduce expression of a broad range of genes *in vivo*, as measured by *lacZ* transcriptional fusions. The biological function of this activity is still unclear (pers. comm. Dr. Lopez de Saro). *YabJ* was selected for further analysis because, like the *hit* family of proteins, it seems to be ubiquitous in nature with both eukaryotic and prokaryotic counterparts. In mouse the YabJ homolog was found to be a heat-responsive protein (EMBL: U50631; HRP12; direct submission to database by Samuel *et al.*). In goat it has been reported to

constitute a tumor antigen UK114 (Ceciliani *et al.*, 1996). In *Lactococcus lactis* it was described as the putative regulator AldR (Godon *et al.*, 1992), and in *Helicobacter pilori* as a putative translation initiation inhibitor (EMBL: G2314082; (Tomb *et al.*, 1997). Dr. Lopez de Saro kindly provided us with the *rpoE* mutant strain HB6010, in which *rpoE* is replaced by the CAT (Cm^r) marker. Chromosomal DNA of this strain was used to transform *B. subtilis* HVH75, and the resulting *hit*$^-$/*rpoE*$^-$ strain was named HVH77. P_{hit}~*lacZ* expression was analysed and compared between strains HVH75 (*hit*$^-$), HVH77 (*hit*$^-$/*rpoE*$^-$) and HVH78 (*hit*$^-$/*yabJ*$^-$). These data are presented in Fig.VII.5.

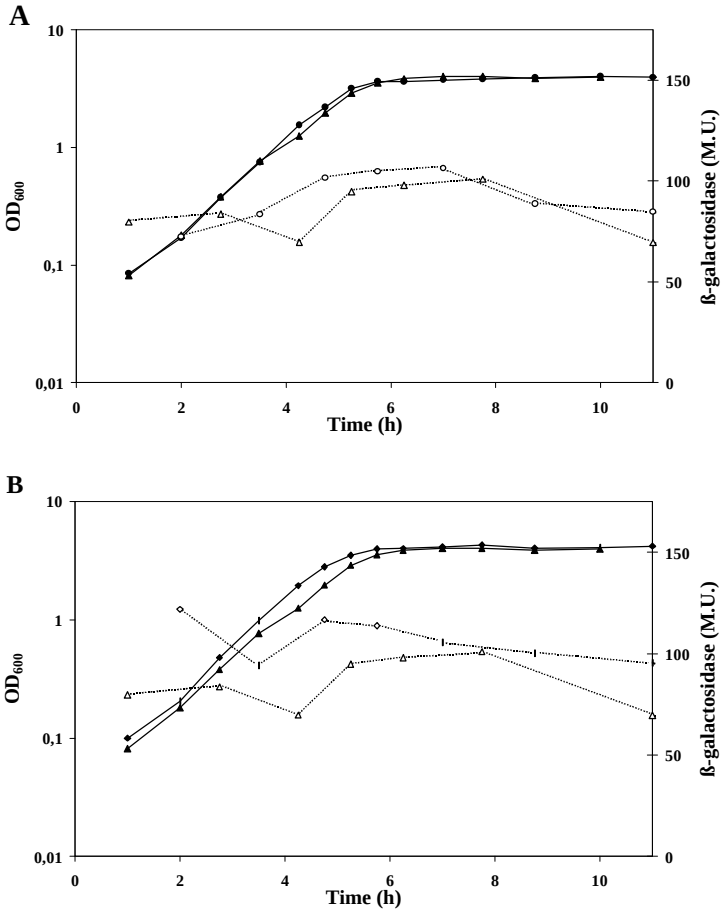


Fig.VII.5. Growth and P*hit*~*lacZ* expression analysis of HVH75 (*hit*$^-$) in comparison with HVH78 (*hit*$^-$/*yabJ*$^-$) (**A**), and HVH77 (*hit*$^-$/*rpoE*$^-$) (**B**). Filled symbols: OD_{600}; open symbols: *lacZ* expression in Miller Units. Triangles: HVH75; circles: HVH78; squares: HVH77.

Only minor differences in *lacZ* expression were observed between the strains. Since *lacZ* transcriptional analysis is a rather insensitive method, we also analysed *hit* transcription in the wild-type strain and in the *yabJ* mutant by Northern analysis. This yielded insight into the transcriptional organisation of the chromosomal region around the *hit* gene as well. The results of these analyses, with cultures grown in sporulation medium (DSM), are depicted in Fig.VII.6. The possible *hit* mRNAs indicated in Fig VII.6B were, except for the 450 nt transcript, not confirmed, but deduced from the Northern results by taking into account the co-ordinates and orientations of genes and terminators in this region.

In the wild-type strain, two transcripts containing the *hit* gene were detected during logarithmic growth. These transcripts are 450 nt and 2600 nt long, respectively. The 450 nt transcript is most abundant, comprising >95% of total *hit* mRNA during logarithmic growth.

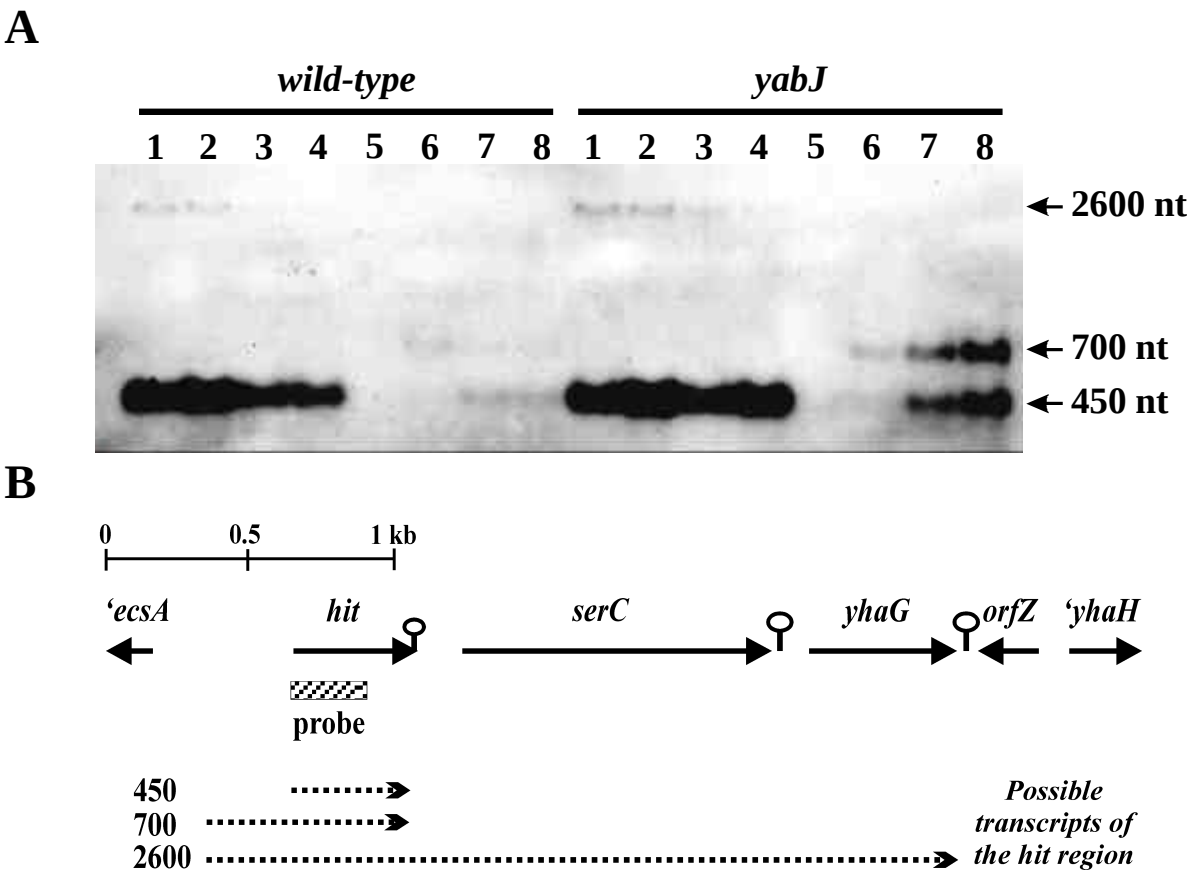


Fig.VII.6. Transcriptional analysis of the *hit* gene in wild-type and *yabJ* mutant background, with cultures grown in DSM. In A, the autoradiogram of the Northern analysis of *hit* mRNA is shown. Numbers above this figure correspond to the time points indicated above the growth-curves presented in C. B: the possible transcriptional organisation of the *hit* chromosomal region. Also indicated here is the probe that was used for the Northern analysis. In C, the growth curves of *B. subtilis* 168 (squares) and the *yabJ* mutant (triangles) are presented. See text for further details.

The 2600 nt transcript was present in small amounts (<5%) during logarithmic growth. In all media tested (sporulation, minimal, and rich medium), the 450 nt and the 2600 nt transcript were present in about the same relative amounts during the logarithmic growth phase (data not shown). In absolute amounts, *hit* mRNA was most abundant in rich and minimal medium and lowest in sporulation medium. This is consistent with the expression data obtained via ß-galactosidase measurements, as described in a previous section. Late in the stationary growth phase, in sporulation medium only, *hit* transcripts were detected again. In *B. subtilis* 168, very low amounts of the 450 nt transcript were observed, and, even less abundant, an additional

transcript of about 700 nt. During this growth phase, the 2600 nt transcript was not detected at all.

In the *yabJ* mutant, the 450 nt and 2600 nt transcripts were present in about the same amounts as in the wild-type during logarithmic growth. Late in the stationary growth phase however, the 450 nt and 700 nt transcripts were much more abundant in the *yabJ* mutant than in the wild-type. This strongly suggests that *yabJ* is a negative regulator of *B. subtilis hit* gene transcription, at least late in the post-exponential growth phase.

The 450 nt transcript probably starts at the adenine residue that is 31 bp upstream of the ATG start codon (primer extension results; data not shown). A possible σ^H –10 sequence is located 7 bp upstream of this start site (GAAT), but Northern analysis of *hit* mRNA in a *sigH* mutant did not confirm the involvement of SigH in *hit* transcription (data not shown). No other putative sigma factor recognition sequences were identified in this region. Instead, we identified several repeated sequence elements in the *hit* promoter region that may be involved in regulation of *hit* transcription (Fig.VII.7). The function of these repeats is unknown, and they are not similar to any known DNA-binding regulator sequences. We have identified a possible rho-independent terminator for this transcript that places the stop codon of the *hit* gene in the loop of the terminator: CCTCATC>A**TAA$_{stop}$**<GATGAGG, with a calculated free energy value of –9.14 kCal.

```
          RR1        R1
TTCCCTTTTAACAAATGATGAATTTGCTGTCATAAAGAGACAGGGTTCTGGACATCATAAAAGTTGTAGA  70

                          R2
GGAAAAGCATATCAACTGAAAAATGAGGTGTTTGTCAAAGACGTCTTAGAAACAGTTGGATAGAAAATAA 140

ACATGTCCAATGCTCAAACCATTTCTCTTCACTGAATTGATGAATTCTTCATGTATGAAGCAAATTGTCA 210

      R3                                        R4                  R5
AATGAGGTGTTTATCGCAGGAATCGTTAGGCATTAAACAAGTCTTCATTTTATTGACAAATGAGGTGCTT 280

ACCAAAGGCATAACACATTTTCTTCATATAAGCTCTTCCTCTGCATTCAGGGTGAACGCTCGCCGTTCAT 350

                R6                                  T     RR2         RBS
CCTGTTTTCTATTTTCTGCATTTCTGTGGTACGATGAATGTATACATACTAAACAATTTCATAAGGAGGA 420

      start
ACCCTCATGCATTGTGCAGAGAATTGTATCTTTTGTAAAATTATTGCCGGCGACATTCCATCAGCGAAGG 490
      M  H  C  A  E  N  C  I  F  C  K  I  I  A  G  D  I  P  S  A  K  V
```

Fig.VII.7. The *hit* promoter region. Indicated are two "reversed" repeats (RR; bold & boxed: TTTAACAAAT) and six (inverted) repeats (R; bold & underlined). The consensus sequence of the repeat is AAATGAGGTGTTT. Only R2 and R3 are perfect, R5 has one difference, R1 has three; R4 & R6 are inverted and less conserved. The confirmed transcriptional start site is indicated with a boxed T. RBS: ribosomal binding site.

Biochemical characterisation of the Hit protein

To investigate the biochemical properties of the *B. subtilis* Hit protein, the carboxy-terminally histidine-tagged protein (Hit~His_6) was incubated with various nucleotidyl compounds: AMP, ADP, ATP, ApppA (Ap_3A), and AppppA (Ap_4A). The Hit protein appeared to be active only with ADP as substrate, and the HPLC analysis of this reaction is presented in Fig.VII.8.

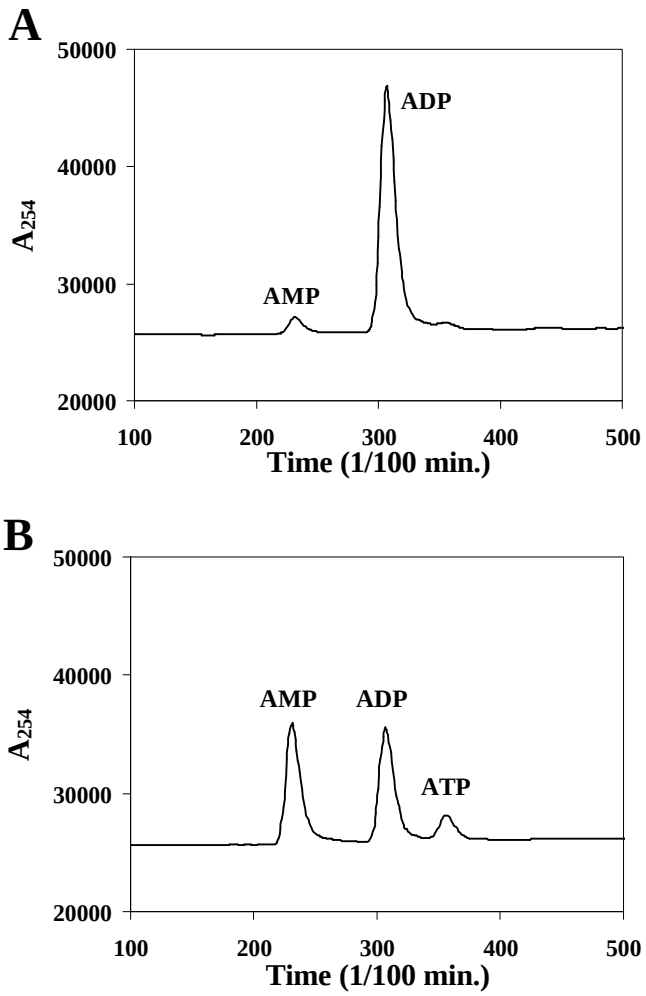


Fig.VII.8. HPLC analysis of the biochemical activity of HIT~His_6 on ADP, at pH 6. In panel **A**, the reaction products of HIT~His_6 after incubation with ADP for 1 h at 0°C are shown. ADP is very slowly degraded to AMP and pyrophosphate (not visible with this detection method) without protein. Panel **B**: the reaction products of HIT~His_6 after incubation with ADP for 1 h at 37°C.

As the hydrolysis of ADP at 37°C in the absence of the Hit protein was essentially the same as represented in Fig.VII.8A, we conclude that, *in vitro*, at pH 6 in the presence of Mn^{2+} ions, the *B. subtilis* Hit protein specifically degrades ADP to AMP and Pi, and to a lesser extent acts as phosphotransferase in the reaction ADP + ADP → ATP + AMP. Similar results were obtained with a hybrid protein in which Hit was amino-terminally fused to the maltose-binding protein (MBP~Hit; data not shown), indicating that the Hit characteristics tested were not affected by either the histidine tag or the fusion to maltose-binding protein. The observed biochemical activity, at least as far as the hydrolysis of ADP is concerned, is in accordance with the findings of Lima *et al.* (1997) with the human PKCI homolog.

To further characterise the Hit protein biochemically, we have analysed in some detail the influence of the pH and the presence of different divalent metal ions on the reaction

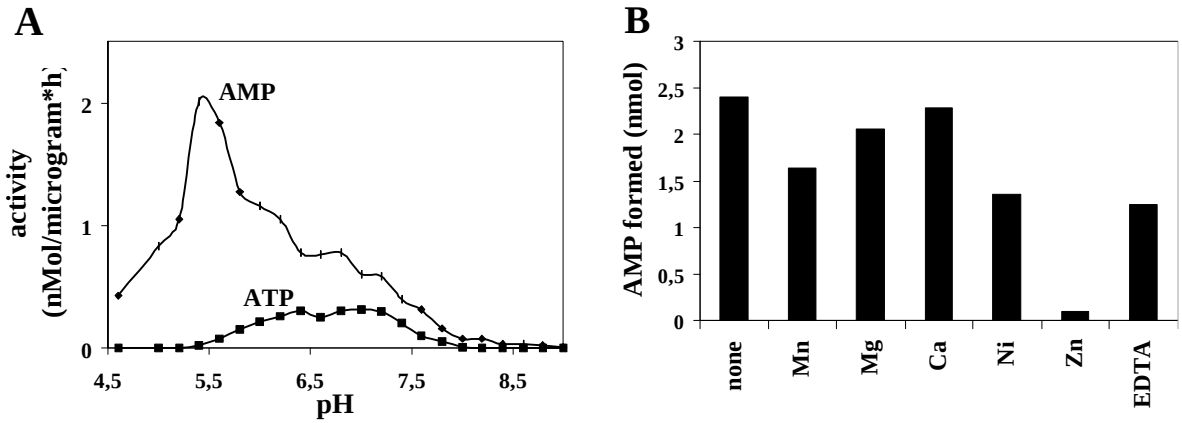


Fig.VII.9. Influence of pH (**A**) and different divalent metal ions (**B**) on the reaction products and reaction velocity of HIT~His_6 with ADP as substrate. In (A), all reactions were done with 80 µM ADP, at 37°C for 1 h. In (B), the reactions were all carried out with 80 µM ADP, at 37°C (pH 5.6, 1 h). Mn^{2+} was substituted with the respective other divalent metal ion or 1 mM of the chelating agent EDTA.

products and the reaction velocity. These data are presented in Fig. VII.9. The relative proportions of AMP and ATP formed were dependent on the pH (Fig. VII.9A). At physiological pH conditions (around pH 7), the *B. subtilis* Hit protein generated from 3 moles of ADP about 1 mole of ATP, 2 moles AMP, and 1 mole of Pi. From the experiment in which we investigated the influence of divalent metal ions, we concluded that zinc has an inhibitory effect on Hit activity. This is an interesting finding, since bovine PKCI-1 and human PKCI-1 have been found to bind specifically zinc *in vitro* (Pearson *et al.*, 1990; Lima *et al.*, 1996). The observation that Hit activity was not strongly influenced by the presence of EDTA suggests that metal ions are not required for catalysis. This is also in accordance with results obtained with human FHIT. Lima *et al* (1997) have reported that the activity of FHIT protein was not much affected in the presence of EDTA. Thus, although it is not necessary for catalysis with ADP as substrate, Zn^{2+} may conceivably inhibit Hit activity *in vivo*.

References

Barnes, L. D., Garrison, P. N., Siprashvili, Z., Guranowski, A., Robinson, A. K., Ingram, S. W., Croce, C. M., Ohta, M., & Huebner, K. (1996). Fhit, a putative tumor suppressor in humans, is a dinucleoside 5',5'''-P^1,P^3-triphosphate hydrolase. *Biochemistry* **35**, 11529-11535.

Brenner, C., Garrison, P., Gilmour, J., Peisach, D., Ringe, D., Petsko, G. A., & Lowenstein, J. M. (1997). Crystal structures of HINT demonstrate that histidine triad proteins are GalT-related nucleotide-binding proteins. *Nat.Struct.Biol.* **4**, 231-238.

Bron, S. (1990). Plasmids. *In* Molecular Biological Methods for *Bacillus*, pp. 75-174. Edited by C. R. Harwood and S. M. Cutting. Chichester: John Wiley & Sons.

Bron, S. & Venema, G. (1972). Ultraviolet inactivation and excision repair in *Bacillus subtilis*. I. Construction and characterization of a eightfold auxotrophic strain and two ultraviolet-sensitive derivatives. *Mutat.Res.* **15**, 1-10.

Bustos, S. A., Schaefer, M. R., & Golden, S. S. (1990). Different and rapid responses of four cyanobacterial *psbA* transcripts to changes in light intensity. *J.Bacteriol.* **172**, 1998-2004.

Ceciliani, F., Faotto, L., Negri, A., Colombo, I., Berra, B., Bartorelli, A., & Ronchi, S. (1996). The primary structure of UK114 tumor antigen. *FEBS Lett.* **393**, 147-150.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A., & Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403.

Fröhlich, K.-U., Fries, H.-W., Rüdiger, M., Erdmann, R., Botstein, D., & Mecke, D. (1991). Yeast cell cycle protein CDC48p shows full-length homology to the mammalian mrotein VCP and is a member of a protein family involved in secretion, peroxisome formation, and gene expression. *J.Cell.Biol.* **114**, 443-453.

Gilmour, J., Liang, N., & Lowenstein, J. M. (1997). Isolation, cloning and characterization of a low-molecular-mass purine nucleoside- and nucleotide-binding protein. *Biochem.J.* **326**, 471-477.

Godon, J. J., Chopin, M. C., & Ehrlich, S. D. (1992). Branched-chain amino acid biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. *J.Bacteriol.* **174**, 6580-6589.

Guan, C., Li, P., Riggs, P. D., & Inouye, H. (1987). Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* **67**, 21-30.

Guérout-Fleury, A.-M., Shazand, K., Frandsen, N., & Stragier, P. (1995). Antibiotic-resistance cassettes for *Bacillus subtilis*. *Gene* **167**, 335-337.

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Huang, Y., Garrison, P. N., & Barnes, L. D. (1995). Cloning of the *Schizosaccharomyces pombe* gene encoding diadenosine 5',5'''-P^1,P^4-tetraphosphate (Ap_4A) asymmetrical hydrolase: sequence similarity with the histidine triad (HIT) protein family. *Biochem.J.* **312**, 925-932.

Inoue, H., Nojima, H., & Okayama, H. (1990). High efficiency transformation of *Escherichia coli* with plasmids. *Gene* **96**, 23-28.

Ish-Horowicz, D. & Burke, F. J. (1981). Rapid and efficient cosmid cloning. *Nucleic Acids Res.* **9**, 2989-2999.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S.-K., Codani, J.-J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Düsterhöft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S. J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, H., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., &

Danchin, A. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Lima, C. D., Klein, M. G., & Hendrickson, W. A. (1997). Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278**, 286-290.

Lima, C. D., Klein, M. G., Weinstein, I. B., & Hendrickson, W. A. (1996). Three-dimensional structure of human protein kinase C interacting protein 1, a member of the HIT family of proteins. *Proc.Natl.Acad.Sci.USA* **93**, 5357-5362.

Maksel, D., Guranowski, A., Ilgoutz, S. C., Moir, A., Blackburn, M. G., & Gayler, K. R. (1998). Cloning and expression of diadenosine 5',5'''-P^1,P^4-tetraphosphate hydrolase from *Lupinus angustifolius* L. *Biochem.J.* **329**, 313-319.

Man, S., Ellis, I. O., Sibbering, M., Blamey, R. W., & Brook, J. D. (1996). High levels of allele loss at the *FHIT* and *ATM* genes in non-comedo ductal carcinoma *in situ* and grade I tubular invasive breast cancers. *Cancer Res.* **56**, 5484-5489.

Mandel, M. & Higa, A. (1970). Calcium-dependent bacteriophage DNA infection. *J.Mol.Biol.* **53**, 159-162.

Martinez, E., Bartolomé, B., & Cruz, F. d. l. (1988). pACYC-derived cloning vectors containing the multiple cloning site and *lacZ_* reporter gene of pUC8/9 and pUC18/19 plasmids. *Gene* **68**, 159-162.

McDonald, J. R., Gröschel-Stewart, U., & Walsh, M. P. (1987). Properties and distribution of the protein inhibitor (M_r 17000) of protein kinase C. *Biochem.J.* **242**, 695-705.

McDonald, J. R. & Walsh, M. P. (1985). Ca^{2+}-binding proteins from bovine brain including a potent inhibitor of protein kinase C. *Biochem.J.* **232**, 559-567.

Miller, J. H. (1982). Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Mozier, N. M., Walsh, M. P., & Pearson, J. D. (1991). Characterization of a novel zinc-binding site of protein kinase C inhibitor-1. *FEBS Lett.* **279**, 14-18.

Negrini, M., Monaco, C., Vorechovsky, I., Ohta, M., Druck, T., Baffa, R., & Huebner, K. (1996). The *FHIT* gene at 3p14.2 is abnormal in breast carcinomas. *Cancer Res.* **56**, 3173-3179.

Noback, M. A., Holsappel, S., Kiewiet, R., Terpstra, P., Wambutt, R., Wedler, H., Venema, G., & Bron, S. (1998). The 172 kb *prkA-addAB* region from 83° to 97° of the *Bacillus subtilis* chromosome contains several dysfunctional genes, the *glyB* marker, many genes encoding transporter proteins, and the ubiquitous *hit* gene. *Microbiol.* **144**, 859-875.

Ohta, M., Inoue, H., Cotticelli, M. R., Kastury, K., Baffa, R., Palazzo, J., Siprashvili, Z., Mori, M., McCue, P., Druck, T., Croce, C. M., & Huebner, K. (1996). The *FHIT* gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell* **84**, 587-597.

Panagopoulos, I., Pandis, N., Thelin, S., Petersson, C., Mertens, F., Borg, A., Kristoffersson, U., Mitelman, F., & Aman, P. (1996). The *FHIT* and *PTPRG* genes are deleted in benign proliferative breast disease associated with familial breast cancer and cytogenetic rearrangements of chromosome band 3p14. *Cancer Res.* **56**, 4871-4875.

Pearson, J. D., DeWald, D. B., Mathews, W. R., Mozier, N. M., Zürcher-Neely, H. A., Heinrikson, R. L., Morris, M. A., McCubbin, W. D., McDonald, J. R., Fraser, E. D., Vogel, H. J., Kay, C. M., & Walsh, M. P. (1990). Amino acid sequence and characterization of a protein inhibitor of protein kinase C. *J.Biol.Chem.* **265**, 4583-4591.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc.Natl.Acad.Sci.USA* **85**, 2444-2448.

Séraphin, B. (1992). The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals. *DNA Seq.* **3**, 177-179.

Shridhar, R., Shridhar, V., Wang, X., Paradee, W., Dugan, M., Sarkar, F., Wilke, C., Glover, T. W., Vaitkevicius, V. K., & Smith, D. I. (1996). Frequent breakpoints in the 3p14.2 fragile site, *FRA3B*, in pancreatic tumors. *Cancer Res.* **56**, 4347-4350.

Simpson, G. G., Clark, G., & Brown, J. W. S. (1994). Isolation of a maize cDNA encoding a protein with extensive similarity to an inhibitor of protein kinase C and a cyanobacterial open reading frame. *Biochim.Biophys.Acta* **1222** , 306-308.

Sozzi, G., Alder, H., Tornielli, S., Corletto, V., Baffa, R., Veronese, M. L., Negrini, M., Pilotti, S., Pierotti, M. A., Huebner, K., & Croce, C. M. (1996b). Aberrant FHIT transcripts in Merkel cell carcinoma. *Cancer Res.* **56**, 2472-2474.

Sozzi, G., Veronese, M. L., Negrini, M., Baffa, R., Cotticelli, M. R., Inoue, H., Tornielli, S., Pilotti, S., De Gregorio, L., Pastorino, U., Pierotti, M. A., Ohta, M., Huebner, K., & Croce, C. M. (1996a). The *FHIT* gene at 3p14.2 is abnormal in lung cancer. *Cell* **85**, 17-26.

Spizizen, I. (1958). Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc.Natl.Acad.Sci.USA* **44**, 1072-1078.

Studier, F. W., Rosenberg, A. H., Dunn, J. J., & Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185**, 60-89.

Thiagalingam, S., Lisitsyn, N. A., Hamaguchi, M., Wigler, M. H., Willson, J. K. V., Markowitz, S. D., Leach, F. S., Kinzler, K. W., & Vogelstein, B. (1996). Evaluation of the *FHIT* gene in colorectal cancers. *Cancer Res.* **56**, 2936-2939.

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H.-P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., FitzGerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547.

Vagner, V., Dervyn, E., & Ehrlich, S. D. (1998). A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiol.* **144**, 3097-3104.

Yanagisawa, K., Kondo, M., Osada, H., Uchida, K., Takagi, K., Masuda, A., Takahashi, T., & Takahashi, T. (1996). Molecular analysis of the *FHIT* gene at 3p14.2 in lung cancer cell lines. *Cancer Res.* **56**, 5579-5582.

CHAPTER VIII

Characterization of *yhcN*, a new forespore-specific gene of *Bacillus subtilis*

VIII.1. Summary

A new *Bacillus subtilis* sporulation-specific gene, *yhcN*, has been identified whose expression is dependent on the forespore specific sigma factor σ^G and to a much lesser extent on σ^F. A translational *yhcN-lacZ* fusion is expressed at a very high level in the forespore, and the protein encoded by *yhcN* was detected in the inner spore membrane. A *yhcN* mutant sporulates normally and *yhcN* spores have identical resistance properties to wild-type spores. However, the outgrowth of *yhcN* spores is slower than that of wild-type spores.

VIII.2. Introduction

The *Bacillus subtilis* genome sequencing project has identified an open reading frame termed *yhcN*, which codes for a small protein with an unusual amino acid composition (Noback *et al*., 1996; EMBL accession no. X96983). This gene is located at 84 degrees on the complete sequence of the *B. subtilis* chromosome, between the *cspB* gene and the glycerol operon *glpPFKD* (Kunst *et al*., 1997), and encodes a 189 residue hypothetical protein which contains 16% asparagine residues. YhcN shows no significant sequence similarity to any protein with known function, although *B. subtilis* genomic sequencing has identified an ORF (*ylaJ* (EMBL sequence accession no. Z97025)) coding for a protein with some sequence homology (20% identity and an additional 24% similarity) to YhcN. YhcN also has slight sequence homology (14% identity and an additional 20% similarity) to one region of a large *E. coli* protein (Swiss Prot P15484) which is essential for the biogenesis of mature CS3 pili (Noback *et al*., 1996). The amino-terminal sequence of YhcN (MFGKKQVLASVLLIPLLMTGCGV-) also exhibits significant sequence similarity to sequences at the amino-termini of membrane anchored lipoproteins, including a positively charged region at the amino-terminus, and a stretch of hydrophobic residues followed by the LMTGC sequence which has some similarity to the consensus sequence for lipoprotein cleavage and modification (LLAGC) (Hayashi and Wu, 1990; Noback *et al*., 1996).

Although the function of YhcN could not be inferred from its amino acid sequence, its high level of asparagine residues is similar to the situation in the γ-type small, acid-soluble

spore protein (SASP) which make up a large percentage of the protein in the dormant spores of *B. subtilis* and other related spore formers (Setlow, 1988). In addition, YhcN contains the amino acid sequence KLEVADE; this sequence is very similar to the sequence KLEIASE found in α/β-type SASP of *B. subtilis* (Setlow, 1988) which is the recognition and cleavage site for the protease (termed GPR) that initiates SASP degradation during spore germination. Thus YhcN may be a distant relative of one of the SASP families.

Both α/β- and γ-type SASP are expressed only during sporulation in the developing forespore and their transcription is initiated by RNA polymerase with a forespore specific sigma factor, σ^G (Setlow, 1988; Haldenwang, 1995). $E\sigma^G$ has a unique promoter specificity, and transcribes a large number of genes expressed only in the forespore - in particular genes that encode much of the protein within the mature spore (Haldenwang, 1995). In this work we have analyzed the expression of *yhcN* and have found it to be expressed only in the forespore of sporulating cells and transcribed almost exclusively by $E\sigma^G$. The *yhcN* gene or gene product also plays a role, either directly or indirectly, in spore outgrowth.

VIII.3. Materials and Methods

Bacterial strains and media

Escherichia coli strain TG1 (Sambrook *et al*., 1989) was used for cloning. The *B. subtilis* strains used in this work are listed in Table VII.1. *B. subtilis* strains with the PS832 background were used to study *yhcN* expression and for analysis of the phenotype of the *yhcN* mutant; *B. subtilis* strains with a PY79 genetic background were used for studies of the genetic dependence of *yhcN* expression. PS832 and PY79 are very similar wild-type strains of *B. subtilis*, but PS832 sporulates more efficiently, while a number of mutations in genes for sporulation sigma factors are available in the PY79 background.

Table VIII.1. Bacterial strains

Bacillus subtilis strains		
Strain	Description	Source (reference)
PS435[a]	*sspE*::*sspE-lacZ*[b] Cm^r	(Mason, J.M. *et al*., 1988)
PS832[a]	wild type derivative of strain 168	laboratory stock
IB331[a]	*yhcN*::*yhcN-lacZ*[b] Cm^r	pIB278↔PS832
IB333[a]	*amyE*::*yhcN-lacZ*[b] Cm^r	pIB320↔PS832
IB345[a]	Δ*yhcN*::*spc* Sp^r	see section *2.6*
IB368[a]	*amyE*::*yhcN-lacZ*[b] Em^r	pCm::Erm (Steinmetz and Richter, 1994)↔IB333
IB373[a]	*amyE*::*yhcN-lacZ*[b] *spoIIIG* Cm^r Em^r	RL560↔IB368
IB375[a]	*amyE*::*yhcN-lacZ*[b] *spoIIIG* [pSDA4] Cm^r Em^r Km^r	pSDA4↔IB373
IB377[a]	*amyE*::*yhcN-lacZ*[b] *spoIIIG* [pDG298] Cm^r Em^r Km^r	pDG298↔IB373
IB385[c]	*yhcN*::*yhcN-lacZ*[b] Cm^r	IB331↔PY79
IB387[c]	*spoIIAC yhcN*::*yhcN-lacZ*[b] Cm^r	IB331↔SC1159
IB389[c]	*spoIIGB yhcN*::*yhcN-lacZ*[b] Cm^r Em^r	IB331↔SC137

IB391[c]	*spoIIIG yhcN*::*yhcN-lacZ*[b] Cm^r Km^r	IB331↔RL831
IB393[c]	*spoIVCB yhcN*::*yhcN-lacZ*[b] Cm^r Em^r	IB331↔SC64
IB415[a]	*amyE*::*yhcN-lacZ*[d] Cm^r	pIB414↔PY79
IB419[a]	*yhcN*::*yhcN-lacZ*[d] Cm^r	pIB417↔PY79
PY79[c]	wild type	(Youngman *et al*., 1984)
RL560[c]	*spoIIIG* Cm^r	R. Losick
RL831[c]	*spoIIIG* Km^r	R. Losick
SC64[c]	*spoIVCB* Em^r	S. Cutting
SC137[c]	*spoIIGB* Em^r	S. Cutting
SC1159	*spoIIAC*	S. Cutting

a: Genetic background is PS832; b: Translational *lacZ* fusion; c: Genetic background is PY79; d: Transcriptional *lacZ* fusion

Construction of *B. subtilis* strains containing a transcriptional *yhcN-lacZ* fusion

A fragment encompassing 147 bp upstream of *yhcN* as well as 25 bp of the *yhcN* coding region was amplified by PCR. The primers used were yhcN5' (5'-CCCAAGCTTCCCTCCTTTGCAGT-GTATTC-3') and yhcN TRN3' (5'-CGGGATCCAAGGACTTGTTTTTTTCCAAAC-3'), which contained extra residues including a *Hin*dIII or *Bam*H1 site at their 5' ends (underlined residues). The PCR product was cut with *Hin*dIII and *Bam*H1 and the resulting fragment cloned into *Hin*dIII and *Bam*H1 cut plasmid pDG268, a vector for construction of transcriptional *lacZ* fusions and their subsequent integration into the *amyE* locus on the *B. subtilis* chromosome (Stragier *et al*., 1988). The resulting plasmid, termed pIB414, was linearized with *Bgl*II and used to transform wild-type *B. subtilis* PS832 to Cm^r. Several Cm^r colonies were picked and checked for the lack of amylase activity as described (Cutting and Vander Horn, 1990). One amylase negative Cm^r colony was chosen, and this strain was called IB415.

To construct a *B. subtilis* strain containing a transcriptional *yhcN-lacZ* fusion at the *yhcN* locus, we cut pIB414 with *Hin*dIII and *Cla*I and cloned the fragment containing the *yhcN* promoter and the 5'-end of *lacZ* between the *Hin*dIII and *Cla*I sites of pTKlac (Kenney and Moran, 1987). The resulting plasmid, termed pIB417, was integrated into the *B. subtilis* PS832 chromosome by a single crossover event, selecting for Cm^r transformants. A Cm^r transformant that contained only one copy of the transcriptional *yhcN-lacZ* fusion at the *yhcN* locus, as determined by Southern blot analysis, was called strain IB419. Chromosomal DNA was isolated from strain IB419 and used to transform *B. subtilis* strain PY79 and its derivatives containing different *spo* mutations to Cm^r.

Construction of *B. subtilis* strains containing a translational *yhcN-lacZ* fusion

A fragment encompassing 147 bp upstream of the *yhcN* ORF as well as 27 bp of the coding region was amplified by PCR. The primers used were: yhcN5' (see previous section) and yhcN3' (5'-CGGGATCCGCAAGGACTTGTTTTTTTCC-3'); the yhcN3' primer had extra residues including a *Bam*H1 site at the 5' end (underlined). The PCR fragment was treated with

T4 DNA polymerase to generate blunt ends, cut with *Bam*H1 and the fragment cloned between the *Sma*I and *Bam*H1 sites of pJF751, a vector for construction of translational *lacZ* fusions (Ferrari *et al*., 1985). The resulting plasmid, termed pIB278, was integrated into the PS832 chromosome by a single crossover event with selection for Cm^r. A Cm^r transformant containing only a single copy of the translational *yhcN-lacZ* fusion at the *yhcN* locus as shown by Southern blot analysis was called strain IB331. Chromosomal DNA was isolated from strain IB331 and used to transform *B. subtilis* strains with PY79 backgrounds to Cm^r.

To construct a *B. subtilis* strain containing the translational *yhcN-lacZ* fusion at the *amyE* locus, we cut pIB278 with *Hin*dIII and *Cla*I and cloned the fragment containing the *yhcN* regulatory region and 5'-end of *lacZ* between the *Hin*dIII and *Cla*I sites of pDG268 (Stragier *et al*., 1988). The resulting plasmid, termed pIB320, was linearized with *Bgl*II and used to transform strain PS832 to Cm^r. Cm^r colonies were picked and checked for the lack of amylase activity as described (Cutting and Vander Horn, 1990). One amylase negative Cm^r colony was chosen and the corresponding strain was called IB333.

Analysis of β-galactosidase activity in sporulating cells, spores and vegetative cells

Sporulation of *B. subtilis* cells was induced at 37ºC by the resuspension method (Sterlini and Mandelstam, 1969). Strains with plasmids carrying genes encoding σ^F or σ^G under control of the IPTG inducible *spac* promoter were grown at 37ºC in 2xYT medium (per liter: 16 g tryptone, 10 g yeast extract, and 5 g NaCl) to an $OD_{600\ nm}$ of 0.25. The culture was then split in half, IPTG added to 2 mM to one half, and incubation continued. β-Galactosidase activity was determined with o-nitrophenyl-β-D-galactopyranoside as described (Nicholson and Setlow, 1990); lysozyme (200 μg/ml) was used for cell permeabilization prior to enzyme assay. To analyze β-galactosidase activity in spores, the spores were first decoated and then treated with lysozyme prior to enzyme assays as described (Nicholson and Setlow, 1990). All β-galactosidase specific activities are expressed in Miller units (Miller, 1972).

Determination of the *yhcN* transcription start site

Total RNA was extracted from sporulating cells of *B. subtilis* strain IB333 3.5 hrs after the beginning of sporulation as described (Moran, 1990). The RNA was used in primer extension reactions at 47ºC with avian myeloblastosis virus reverse transcriptase (Moran, 1990) and as primers either yhcN-80 (5'-GCAGCCAGTCATAAGCAAAGG-3'), which anneals only to *yhcN* mRNA, or lacZ-70 (5'-AAGGCGATTAAGTTGGGTAACG-3'), which anneals only to *yhcN-lacZ* mRNA. DNA size standards for analysis of the primer extension products were produced using the same two primers in DNA sequencing reactions. The yhcN-80 primer was used with plasmid pIB321 carrying a 900 bp fragment encompassing the *yhcN* region (see next section), and the lacZ-70 primer with plasmid pIB320 which carries the translational *yhcN-lacZ* fusion in plasmid pDG268.

Construction of the *yhcN* null mutant

Plasmid pSGMU2 (Perego, 1993) was linearized with *Eco*R1, treated with T4 DNA polymerase to fill the ends, and ligated to give plasmid pIB297 which is pSGMU2 cured of its *Eco*R1 site. A 900 bp fragment containing *yhcN* and its flanking regions (Fig. VIII.1.) was amplified by PCR. The primers used were: yhcN_mut_new5' (5'-CCCAAGCTTGATCCTATGATCAATGCTG-3') and yhcN_mut_new3' (5'-CGGGATCCGCAAACGTCTCCCTCC-3'), each containing extra residues including a *Bam*H1 or *Hin*dIII site at their 5' ends (underlined residues). The PCR fragment was cut with *Hin*dIII and *Bam*H1 and cloned between the *Bam*H1 and *Hin*dIII sites of plasmid pIB297 to yield plasmid pIB321. To delete most of *yhcN's* coding region from plasmid pIB321 the plasmid was cut with *Bst*E2, treated with T4 DNA polymerase to fill the ends, and then cut with *Eco*R1. The 4.1 kb fragment was isolated and ligated with a 1.1 kb fragment containing a spectinomycin cassette obtained by digestion of plasmid pJL74 (LeDeaux and Grossman, 1995) with *Eco*R1 and *Ecl*136II. The resulting plasmid, termed pIB325, contains *yhcN* flanking regions with the spectinomycin cassette replacing much of *yhcN*'s coding sequence, and was linearized with *Hin*dIII and used to transform *B. subtilis* PS832 to Sp^r. In this transformation the spectinomycin cassette was integrated into the *B. subtilis* chromosome by a double crossover event removing most of the *yhcN* coding sequence. The expected chromosomal structure of one Sp^r transformant (termed strain IB345) was confirmed by Southern blot analysis.

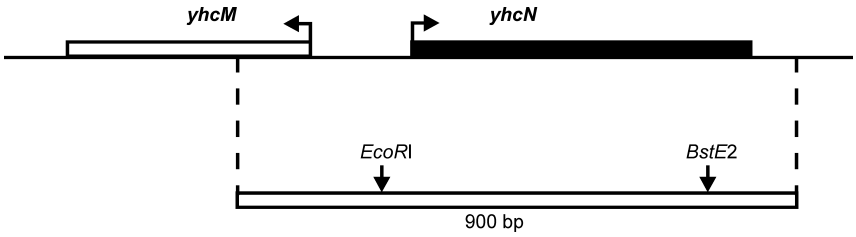


Fig. VIII.1. Physical map of the *yhcM/N* region. The 900 bp fragment used in the construction of the *yhcN* null mutant is shown schematically. The horizontal arrows denote the direction of transcription as shown (*yhcN*, this work) or inferred (*yhcM*). The boxes denote the limits of the *yhcM* and *yhcN* ORFs.

Analysis of resistance, germination and outgrowth of *B. subtilis* spores

Spores were harvested from cultures grown for 48 hours at 37ºC in 2xSG medium (Goldrick and Setlow, 1983; Leighton and Doi, 1971) and purified as described (Mason and Setlow, 1986). Spores in water were heat treated or UV irradiated with 254 nm light and survival measured as described (Mason and Setlow, 1986; Setlow and Setlow, 1987; Fairhead *et al.*, 1993).

Purified spores in water were heat activated for 30 min at 70ºC, cooled on ice, and then diluted to an $OD_{600\ nm}$ of around 0.4 in 2xYT medium containing 4 mM L-alanine, or to an

$OD_{600\ nm}$ of around 0.8 in Spizizen's minimal medium (Spizizen, 1958) without Casamino acids but containing 4 mM L-alanine and 50 μg/ml L-tryptophan. Cultures were incubated at 37ºC with good aeration and the $OD_{600\ nm}$ of the cultures was monitored.

Analysis of proteins from *B. subtilis* spores and spore membranes

About 125 $OD_{600\ nm}$ units of *B. subtilis* spores were lyophilized and dry-ruptured (8 times for 1 min each) with a dental amalagamator (Wig-L-Bug) using glass beads (~ 100 mg) as the abrasive. The dry powder was extracted with 0.6-0.8 ml of ice cold buffer (25 mM Tris-HCl (pH 7.4), 5 mM EDTA, 0.3 mM PMSF) for 30 min on ice, followed by centrifugation for 2 min in a microcentrifuge. Aliquots of both the supernatant and pellet fractions were analyzed by SDS-PAGE on a 15% gel.

Spore coats were removed from cleaned spores by incubation in 0.1 M DTT, 0.1 M NaCl, 0.1 M NaOH, 0.5% SDS (pH 10) at 65ºC for 30 min. This procedure removes much spore coat protein as well as the spore's outer membrane, and allows subsequent rupture of spores with lysozyme. The decoated spores were washed extensively (> 10 times) with distilled water and resuspended in buffer A (50 mM Tris-HCl (pH 7.4), 5 mM EDTA, 1 mM PMSF, 0.2 mg/ml $MgCl_2$, 1 μg/ml DNAse I) at 50-75 OD 600 U/ml. Lysozyme was added to a final concentration of 1 mg/ml and the mixture was incubated for 2 min at 25ºC and then for 20 min on ice. The lysate was sonicated for 1 min with glass beads present to shear the spore's inner membrane from the cortex and germ cell wall, and the resulting extract was centrifuged in a microfuge for 5 min. The supernatant fluid was centrifuged at 28,000 x g for 5 min to pellet spore cortex and cell wall fragments, and the membrane fraction was recovered from the supernatant fluid by centrifugation at 100,000 x g for 1 hr. The membrane pellet was washed with buffer A and resuspended in 50 μl buffer A. Membranes from vegetatively growing cells (OD ≅ 1.0 in 2xSG medium) were recovered by an essentially identical protocol except that the cells were not treated with decoating solution. Approximately 10 μg of membrane protein were run on SDS-PAGE using a Tris/Tricine polyacrylamide gel (Schagger and Von Jagow, 1987), proteins transferred to polyvinylidene difluoride paper, and the paper stained with Coomassie Blue. The 22 kDa band tentatively identified as YhcN was digested with trypsin, peptides purified by high pressure liquid chromatography, and a peptide sequenced as described previously (Patel-King *et al*., 1996).

Synthesis of YhcN in *E. coli* and its cleavage with GPR

In order to express *yhcN* in *E. coli*, the *yhcN* ORF was amplified by PCR. The primers used were 5'pET (5'-GGAATTCCATATGTTTGGAAAAAAACAAGTCC-3') and 3'pET (5'-CCGCTCGA-GTTCAGCGTTAGGGAATACAC-3'), each of which had extra residues at their 5' ends (underlined) including *Nde*I and *Xho*I sites. The PCR product was cleaved with *Nde*I and *Xho*I, and cloned between the *Nde*I and *Xho*I sites of the expression vector

pET29(b+) (Novagen, Milwaukee, WI). The resulting plasmid, termed pIB410, was introduced into the *E. coli* expression strain BL21(DE3) (Novagen, Milwaukee, WI) and *yhcN* expression was induced with IPTG according to the Novagene pET System Manual. 1.5 hours after induction cells were collected, disrupted with lysozyme, crude soluble and insoluble fractions prepared according to the pET System Manual, and aliquots analyzed by SDS-PAGE along with aliquots from uninduced cells. The insoluble fraction of induced cells had one new prominent protein band migrating at 22 kDa, which is close to the size expected for YhcN (21 kDa). The insoluble fraction of induced cells was lyophilized, weighed and dissolved in 8 M urea - 25 mM Tris-HCl (pH 7.5) for 1 h at room temperature to obtain a protein concentration of ~ 10 mg/ml.

To test for GPR cleavage of YhcN produced in *E. coli*, 300 µl of a mix containing 0.8 M urea, 5 mM Tris-HCl (pH 7.5), 2.5 mM $CaCl_2$, and 300 µg of the protein from the induced cells with or without 20 µg of active *B. megaterium* GPR purified as described (Illades-Aguiar and Setlow, 1994) was incubated for 30 min at 37ºC. The YhcN remained soluble during this incubation. To check that GPR was active in this mixture we also carried out the same reaction with the GPR substrate SspC (60 µg) (Loshon *et al*., 1997). Aliquots of reaction mixes were analyzed on SDS-PAGE for YhcN and on a polyacrylamide gel run at low pH (Illades-Aguiar and Setlow, 1994) for SspC.

VIII.4. Results

Expression of *yhcN* during growth and sporulation

To examine *yhcN* expression we constructed transcriptional and translational *yhcN-lacZ* fusions, placed them at both the *yhcN* and *amyE* loci, and measured *yhcN*-directed β-galactosidase activity during vegetative growth and sporulation, and in dormant spores. No significant expression of the *yhcN-lacZ* fusions was observed in vegetatively growing cells (Fig. VIII.2.).

However, the transcriptional and translational *yhcN-lacZ* fusions were both expressed beginning ~ 2 hours after induction of sporulation, with the maximum β-galactosidase specific activity after 4-5 hours (Fig. VII.2.). Surprisingly, the translational *yhcN-lacZ* fusion was expressed to a much higher level than the transcriptional fusion, although both fusions exhibited similar kinetics of expression during sporulation (Fig. VII.2A,B.). Analysis of the β-galactosidase level in spores of the different *yhcN-lacZ* fusions showed that more than 75% of the *yhcN*-driven β-galactosidase accumulated during sporulation was incorporated into the mature spore (data not shown). These data indicate that *yhcN* is not only a sporulation gene, but is also a forespore specific gene. The similar kinetics and level of β-galactosidase expression from the transcriptional *yhcN-lacZ* fusion incorporated at the *yhcN* and *amyE* loci (Fig. VIII.2A.) further indicate that the 147 bp upstream of *yhcN* that were inserted at *amyE* likely contain the complete *yhcN* promoter. Similar results were obtained with the translational *yhcN-lacZ* fusion (Fig. VIII.2B.).

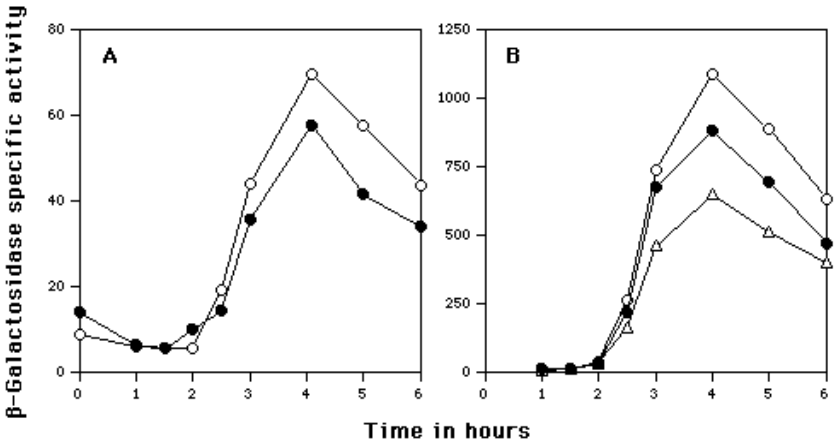


Fig. VIII.2. Expression of transcriptional (A) or translational (B) *yhcN-lacZ* fusions during sporulation. Cells of various strains were sporulated by the resuspension method, and samples taken and assayed for β-galactosidase. Time zero is the time of resuspension of the culture to initiate sporula-tion. The symbols used for the various strains are: A) O, IB415 (*amyE*::*yhcN-lacZ*); and ●, IB419 (*yhcN*::*yhcN-lacZ*); B) O, IB333 (*amyE*::*yhcN-lacZ*); ●, IB331 (*yhcN*::*yhcN-lacZ*); and ∆, PS435 (*sspE*::*sspE-lacZ*).

Sigma factor dependence of *yhcN* expression

The timing of *yhcN-lacZ* expression was essentially identical to that of *sspE*, another forespore specific gene (Fig. VIII.2B.). Since *sspE* is known to be transcribed almost exclusively by RNA polymerase containing the forespore specific sigma factor σ^G (Sun *et al.*, 1989), this suggests that *yhcN* may also be under σ^G control. To prove the σ^G-dependence of *yhcN* expression we analyzed the expression of the translational *yhcN-lacZ* fusion in different *spo* mutant backgrounds. A mutation in the *spoIVCB* gene coding for a part of the late mother cell specific sigma factor σ^K had no significant effect on *yhcN-lacZ* expression (Fig. VIII.3A.).

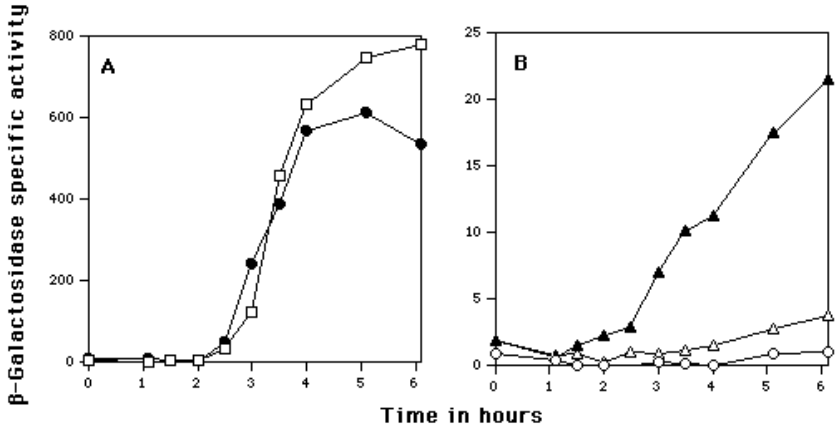


Fig. VIII.3. Expression of the translational *yhcN-lacZ* fusion in various *spo* mutants. Strains with a PY79 background were sporulated and β-galactosidase was assayed as described in Methods. Time zero is the time of resuspension of the culture to initiate sporulation. The symbols used for the various strains are: ●, IB385 (spo$^+$); O, IB387 (*spoIIAC*); ▲, IB389 (*spoIIGB*); ∆, IB391 (*spoIIIG*); and ð, IB393 (*spoIVCB*). Note the different scales in A and B.

In contrast, a mutation in *spoIIIG* which codes for the late forespore-specific sigma factor σ^G decreased the level of *yhcN*-driven *lacZ* expression to only ~ 4% that of the wild type level (Fig. VIII.3B.). However, a mutation in the *spoIIAC* gene which codes for the early forespore-specific sigma factor σ^F abolished *yhcN-lacZ* expression (Fig. VIII.3B.). Since σ^F is required for synthesis of σ^G (Haldenwang, 1995), these data indicate that *yhcN* is a forespore-specific gene which is transcribed primarily by Eσ^G and to a small extent by Eσ^F. *yhcN-lacZ* expression in a *spoIIGB* mutant lacking the mother cell specific σ^E was higher than in a σ^G mutant (Fig. VIII.3B.), as was observed previously for other genes fully or partially dependent on σ^F (Lewis *et al*., 1994; Karow *et al*., 1995; Londono-Vallejo, Stragier, 1995; Londono-Vallejo *et al*., 1997). Similar results were obtained when the sigma factor dependence of the expression of the transcriptional *yhcN-lacZ* fusion during sporulation was analyzed (data not shown).

To prove conclusively that σ^F is able to direct the transcription of *yhcN*, we introduced plasmid pSDA4 (Shazand *et al*., 1995), which contains the structural gene for σ^F under the control of the IPTG-inducible *spac* promoter, into a strain containing a translational *yhcN-lacZ* fusion as well as a mutation in *spoIIIG*. Upon induction of σ^F synthesis in vegetatively growing cells of this strain we obtained some increase in β-galactosidase activity (Fig. VIII.4A.), showing that σ^F was able to direct a low level of expression of *yhcN-lacZ*. However, vegetative cells containing plasmid pDG298 (Sun *et al*., 1989) carrying *spac-spoIIIG* showed more than 100 fold higher expression of the *yhcN-lacZ* fusion upon induction of σ^G synthesis (Fig. VIII.4B.). Therefore, we conclude that *yhcN* is transcribed primarily by Eσ^G but can also be recognized to a small extent by σ^F. However, it is unclear whether this σ^F dependent expression

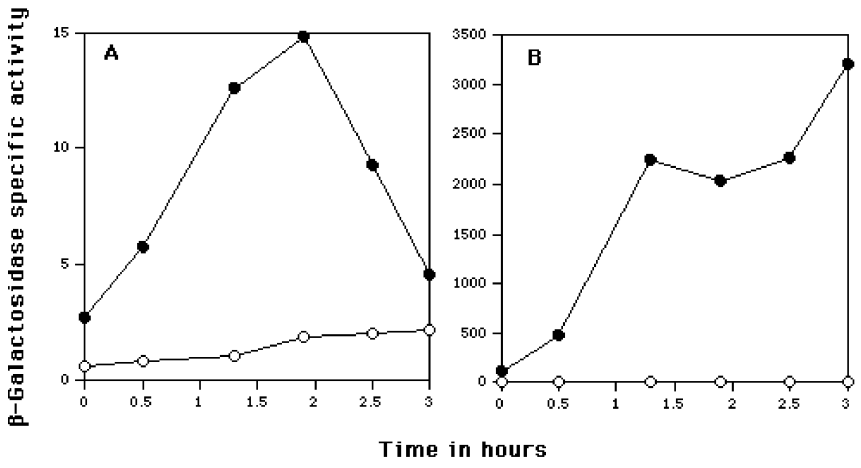


Fig. VIII.4. Induction of *yhcN-lacZ* expression in vegetatively growing cells engineered to produce A) σ^F or B) σ^G. Cells were grown at 37°C in 2xYT medium, and at an OD_{600}=0.25 cultures were split in half, one half made 2 mM in IPTG, and samples taken from both cultures for assay of β-galactosidase. The strains and symbols used are: A) IB375 (P*spac*-σ^F; B) IB377 (P*spac*-σ^G); O - without IPTG; and ● - with IPTG.

Localization of the yhcN promoter

To precisely localize the *yhcN* promoter we carried out primer extension analysis using RNA from sporulating cells of strain IB333 containing the translational *yhcN-lacZ* fusion at the *amyE* locus. Two different primers were used for this analysis; one annealed to the *lacZ* portion of *yhcN-lacZ* mRNA while the other annealed only to the *yhcN* mRNA. Both primers gave the same start site for transcription from the *yhcN* promoter (Fig. VIII.5A,B.). However,

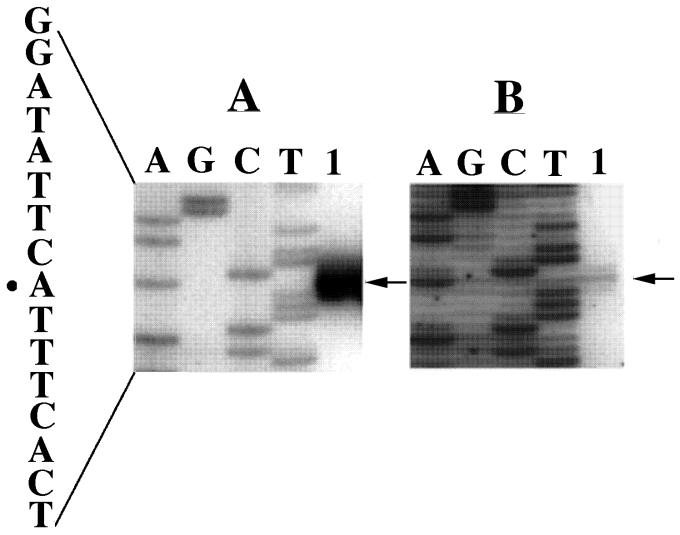


Fig. VIII.5. Primer extension analysis of the start of transcription of *yhcN*. RNA from sporulating cells of strain IB333 was isolated, and primer extension products were obtained and analyzed as described in Methods. The primer used in A is lacZ-70 which anneals only to *lacZ* sequences; the primer used in B is yhcN-80 which anneals only to *yhcN* sequences. The lanes labeled A, G, C and T are DNA sequencing reactions with appropriate primers and either A) pIB320 or B) pIB321. Lanes labeled 1 are primer extension reactions with sporulating cell RNA. Primer extension products are marked with arrows, and the transcription start site on the *yhcN* upstream se-quence to the left of the figure is marked with a dot. The upstream sequence of *yhcN* is taken from Noback, M.A., *et al*., 1996. The intensity of the DNA sequencing lanes in B was approximately equal to that in A when both samples were exposed to film for equivalent amounts of time.

the *yhcN-lacZ* mRNA appeared to be more abundant than *yhcN* mRNA, since we obtained > 10 fold more extension product with the primer annealing to *yhcN-lacZ* mRNA (Fig. VIII.5A.) than with the primer annealing only to *yhcN* mRNA (Fig. VIII.5B.). Transcription of *yhcN* starts 24 nt upstream of the *yhcN* AUG codon, at a T residue (Fig. VIII.5,6.). Sequences centered 10 and 35 nt upstream of the transcription start site show good similarity to the -10 and -35 consensus sequences recognized by both σ^G and σ^F with appropriate spacing (17 nt) between these consensus sequences (Fig. VIII.6.) (Haldenwang, 1995). The *yhcN* sequence has one G residue 3 residues upstream of the -10 region, a feature of several promoters recognized by both σ^F and σ^G. In contrast, G residues both 3 and 2 residues upstream of the -10 sequence are hallmarks of good σ^F dependent promoters (Sun *et al*., 1991).

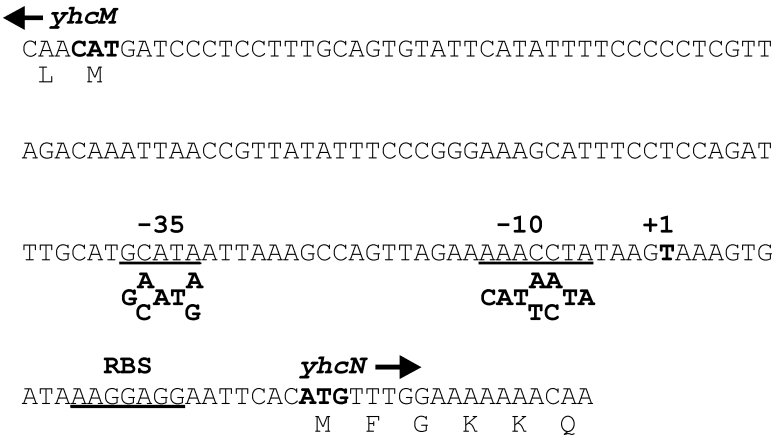


Fig. VIII.6. Sequence of *yhcN* 5' region. Sequences corresponding to -10 and -35 promoter elements and *yhcN* ribosome binding site (RBS) are underlined. Consensus sequences for -10 and -35 promoter elements (Haldenwang, 1995) of σ^G dependent promoters are shown in bold letters, as is the transcription start site.

Analysis of YhcN in *B. subtilis* and its overexpression in *E. coli*

The forespore specific expression of *yhcN*, as well as the capture of most *yhcN* driven β-galactosidase in the dormant spore, suggested that YhcN is a spore protein, although SDS-PAGE of aliquots of total soluble and insoluble protein from disrupted spores did not reveal any significant difference in the protein pattern from wild-type and *yhcN* mutant spores (data not shown). However, SDS-PAGE of proteins from vegetative cell membranes and the inner spore membrane identified a protein of ~ 22 kDa (approximately the predicted size of YhcN

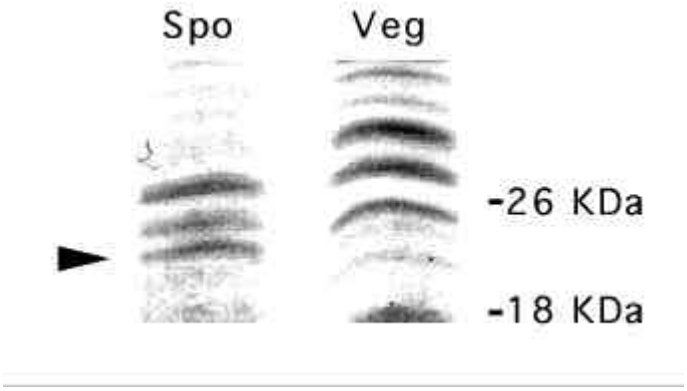


Fig. VIII.7. Analysis of membrane proteins from spores and vegetative cells of *B. subtilis*. The inner spore membrane and vegetative cell membranes were isolated, and ~ 10 μg of protein was run on Tris/Tricine SDS-PAGE as described in Methods, and the gel stained with Coomassie Blue. The samples in the various lanes are from: Spo - spores; and Veg – vegetative cells. The migration position of molecular weight markers is given to the right of the Fig., and the arrowhead to the left points out a 22 kDa protein unique to spores that was shown to be YhcN.

(21 kDa)) that was present in spores but not growing cells (Fig. VIII.7.). Tryptic digestion of this 22 kD protein followed by amino acid sequence analysis of one tryptic peptide gave the sequence NIDNVYVSAN, which is identical to residues 141-150 in YhcN (Noback *et al*., 1996). These data indicate that YhcN is a spore specific membrane protein. However, we estimate that YhcN comprises < 0.1% of the protein of dormant spores. In contrast the α/β-type and γ-type SASP together comprise ~ 12% of total spore protein (Setlow, 1988).

We also overexpressed YhcN in *E. coli*. Although the protein was insoluble it was not in *E. coli* membranes (data not shown). We then tested the crude YhcN as a substrate for cleavage by GPR which initiates cleavage of SASP during spore germination (Setlow, 1988).

Under conditions where cleavage of a good GPR substrate (SspC) was observed, we saw no cleavage of YhcN (data not shown); the rate of cleavage of YhcN by GPR was at least 500 fold slower than cleavage of SspC (data not shown).

Characterization of the *yhcN* null mutant

In addition to using the *yhcN* null mutant to assess the level of YhcN in spores, we also analyzed the properties of this mutant to elucidate a role for *yhcN*. The mutant strain sporulated normally, with kinetics and yield of phase-bright spores identical to those of the wild type. The *yhcN* spores also had the same resistance to heat and UV radiation as did the wild type spores (data not shown). The *yhcN* spores exhibited no defect in the initiation of spore germination, as measured by the initial fall in optical density of a spore culture following mixing of spores with germinant (Fig. VIII.8.). However, *yhcN* spores did have a slight defect in spore outgrowth, as they returned to vegetative growth more slowly than wild type spores (Fig. VIII.8.). This defect was more pronounced when spore germination and outgrowth was in a minimal medium as compared to a rich medium (Fig. VIII.8A,B.).

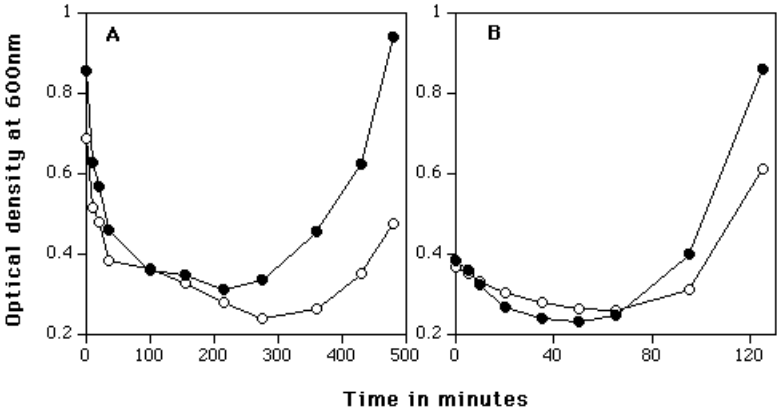


Fig. VIII.8. Germination and outgrowth of spores of the *yhcN* mutation in A) a minimal or B) a rich medium. Spores of strains PS832 (wild type) (●) and IB345 (Δ*yhcN*::*spc*) (O) were germinated in either A) a minimal medium (Spizizen's) or B) a rich medium (2xYT) as described in Methods, and the OD_{600} followed.

VIII.5. Discussion

We initially chose to analyze the expression of *yhcN* because its size, amino acid composition and KLEVADE sequence suggested that YhcN might be a SASP that was susceptible to cleavage by GPR. *yhcN* is expressed only in the developing spore, as are genes coding for both α/β-type and γ-type SASP. However, these SASP are major spore proteins, and levels of YhcN are extremely low in spores; YhcN also does not have other conserved amino acid sequences found in α/β- and γ-type SASP (Setlow, 1988). In addition, YhcN that had been overexpressed in *E. coli* was not cleaved by GPR *in vitro*. The lack of cleavage of YhcN by GPR was somewhat surprising, as this protease tolerates a variety of substitutions in its canonical recognition and cleavage site (EIASE) including substitution of Asp for the first Glu residue, a variety of large hydrophobic residues for Ile, and even Gln for the second Glu (Carillo-Martinez and Setlow, 1994). However, while proteins with these various substitutions are cleaved appropriately by GPR, the cleavage rate can be decreased by more than 10^4. The

conserved Ser residue in the GPR cleavage site can also be varied, as Ala, Arg, Asn, Gln, Leu, Lys and Val have been found in this position, although again at least some of these subsitutions greatly reduce the rate of GPR cleavage (Cabrera-Martinez and Setlow, 1991; Carillo-Martinez and Setlow, 1994; Setlow, 1988). The fact that cleavage of YhcN was > 500 fold slower than that of a good GPR substrate suggests that GPR does not tolerate an acidic residue at the position normally occupied by Ser.

One of the surprising results from this work was the large (> 10 fold) difference in the level of expression of transcriptional and translational *lacZ* fusions to *yhcN*, with the translational fusion giving much higher levels of expression. While the translational *yhcN-lacZ* fusion utilized the same transcription start site as *yhcN* mRNA, we found much more RNA from the translational *yhcN-lacZ* fusion. These data suggest that the mRNA from the translational *yhcN-lacZ* fusion is much more stable than either *yhcN* mRNA or mRNA from the transcriptional *yhcN-lacZ* fusion. An alternative possibility is that since a stop codon in-frame with *yhcN* was generated in the transcriptional *yhcN-lacZ* fusion junction, this stop codon may exert a polar effect on *lacZ* expression. However, we have not studied these points further.

Some of our findings with YhcN are similar to those made previously with SspF. The latter protein is also encoded by a forespore expressed gene; its mRNA is quite abundant, yet SspF is present at minute levels at best in spores (Loshon *et al*., 1997; Ollington and Losick, 1981; Stephens *et al*., 1984). SspF shares weak sequence similarity with α/β-type SASP, and contains the potential GPR cleavage site ELAKD that is cleaved by GPR *in vitro* (Loshon *et al*., 1997; Setlow, 1993). However, mutation of the *sspF* gene had no effect on sporulation, spore resistance, spore germination or spore outgrowth (Loshon *et al*., 1997).

The *yhcN* mutation also had no effect on sporulation, spore resistance or germination, but did retard spore outgrowth. This latter phenotype seems likely to be due to the loss of *yhcN* and not a polar effect on a downstream gene, as *yhcN* has a strong transcription terminator just downstream of the coding region, and the next open reading frame (*yhcO*) is 179 nt downstream of *yhcN* (Noback *et al*., 1996). However, the reason for the effect of the *yhcN* mutation on spore outgrowth is not clear. We have found that YhcN is in the inner spore membrane, and its membrane location is consistent with the similarity between the YhcN amino terminal sequence and sequences at the amino-termini of membrane anchored lipoproteins. However, we do not know if YhcN is a lipoprotein, nor how its membrane location might affect any function of this protein during spore outgrowth.

References

Cabrera-Martinez, R.-M., Setlow, P. (1991). Cloning and nucleotide sequence of three genes coding for small, acid-soluble proteins of *Clostridium perfringens* spores. *FEMS Microbiol. Lett* **77,** 127-132.

Carillo-Martinez, Y., Setlow, P. (1994). Properties of *Bacillus subtilis* small, acid-soluble spore proteins with changes in the sequence recognized by their specific protease. *J. Bacteriol*. **176,** 5357-5363.

Cutting, S.M., Vander Horn, P.B. (1990). Genetic analysis. *In*: Harwood, C.R., Cutting, S.M. (Eds.), Molecular biological methods for *Bacillus*. John Wiley & Sons, Chichester, pp. 27-74.

Fairhead, H., Setlow, B., Setlow, P. (1993). Prevention of DNA damage in spores and *in vitro* by small, acid-soluble proteins from *Bacillus* species. *J. Bacteriol*. **175,** 1367-1374.

Ferrari, E., Howard, S.M.H., Hoch, J.A. (1985). Effect of sporulation mutations on subtilisin expression, assayed using a subtilisin-β-galactosidase gene fusion. *In*: Hoch, J.A., Setlow, P. (Eds.), Molecular biology of microbial differentiation. American Society for Microbiology, Washington, DC, pp. 180-184.

Goldrick, S., Setlow, P. (1983). Expression of a *Bacillus megaterium* sporulation-specific gene in *Bacillus subtilis*. *J. Bacteriol*. **155,** 1459-1462.

Haldenwang, W.G. (1995). The sigma factors of *Bacillus subtilis. Microbiol. Rev*. **59,** 1-30.

Hayashi, S., Wu, H.C. (1990). Lipoproteins in bacteria *J. Bioenerg. Biomemb*. **22,** 451-471.

Illades-Aguiar, B., Setlow, P. (1994). Studies of the protease which initiates degradation of small, acid-soluble proteins during germination of spores of *Bacillus* species. *J. Bacteriol.* **176,** 2788-2795.

Karow, M.L., Glaser, P., Piggot, P.J. (1995). Identification of a gene, *spoIIR*, that links the activation of σ^E to the transcriptional activity of σ^F during sporulation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **92,** 2012-2016.

Kenney, T.J., Moran, C.B. (1987). Organization and regulation of an operon that encodes a sporulation-essential sigma factor in *Bacillus subtilis*. *J. Bacteriol.* **169,** 3329-3339.

Kunst, F., Ogasawara, N., and others. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390,** 249-256.

LeDeaux, J.R., Grossman, A.D. (1995). Isolation and characterization of KinC, a gene that encodes a sensor kinase homologous to the sporulation sensor kinases KinA and KinB in *Bacillus subtilis*. *J. Bacteriol*. **177,** 166-175.

Leighton, T.J., Doi, R.H. (1971). The stability of messenger ribonucleic acid during sporulation in *Bacillus subtilis*. *J. Biol. Chem.* **246,** 3189-3195.

Lewis, P.J., Partridge, S.R., Errington, J. (1994). σ factors, asymmetry, and the determination of cell fate in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **91,** 3849-3853.

Londono-Vallejo, J.-A., Frehel, C., Stragier, P. (1997). *spoIIQ*, a forespore-expressed gene required for engulfment in *Bacillus subtilis*. *Mol. Microbiol*. **24,** 29-39.

Londono-Vallejo, J.-A., Stragier, P. (1995). Cell-cell signalling pathway activating a developmental transcription factor in *Bacillus subtilis*. *Genes Dev.* **9,** 503-508.

Loshon, C.A., Kraus, P., Setlow, B., Setlow, P. (1997). Effect of inactivation or overexpression of the *sspF* gene on properties of *Bacillus subtilis* spores. *J. Bacteriol*. **179,** 272-275.

Mason, J.M., Hackett, R.H., Setlow, P. (1988). Studies on the regulation of expression of genes coding for small, acid-soluble proteins of *Bacillus subtilis* spores using *lacZ* gene fusions. *J. Bacteriol*. **170,** 239-244.

Mason, J.M., Setlow, P. (1986). Essential role of small, acid-soluble spore proteins in the resistance of *Bacillus subtilis* spores to UV light. *J. Bacteriol*. **167,** 174-178.

Miller, J.H. (1972). Measuring gene expression in *Bacillus*, *In*: Harwood, C.R., Cutting, S.M. (Eds.)., Molecular biological methods for *Bacillus*. John Wiley and Sons, Chichester, England, pp. 267-293.

Moran, C.P. (1990). Measuring gene expression in *Bacillus*, *In*: Harwood, C.R., Cutting, S.M. (Eds.), Molecular biological methods for *Bacillus*. John Wiley and Sons, Chichester, England, pp. 267-293.

Nicholson, W.L., Setlow, P. (1990). Sporulation, germination and outgrowth. *In*: Harwood, C.R., Cutting, S.M. (Eds.), Molecular biological methods for *Bacillus*. John Wiley and Sons, Chichester, England, pp. 391-450.

Noback, M.A., Terpstra, P., Holsappel, S., Venema, G., Bron, S. (1996). A 22 kb DNA sequence in the *cspB-glpPFKD* region at 75 degrees on the *Bacillus subtilis* chromosome. *Microbiology* **142,** 3021-3026.

Ollington, J.F., Losick, R. (1981). A cloned gene that is turned on at an intermediate stage of spore formation. *J. Bacteriol.* **147,** 443-551.

Patel-King, R., Benashski, S.E., Harrison, A., King. S.M. (1996). Two functional thioredoxins containing redox-sensitive vicinal dithiols from the clamydomonas outer dynein arm. *J. Biol. Chem*. **271,** 6283-6291.

Perego, M. (1993). Integrational vectors for genetic manipulation in *Bacillus subtilis*. *In*: Sonenshein, A.L., Hoch, J.A., Losick, R. (Eds.), *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology, and molecular genetics. American Society for Microbiology, Washington, DC, pp. 615-624.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989). Molecular cloning: A laboratory manual. 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schagger, H., von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1-100 kDa. *Anal. Biochem*. **166,** 368-379.

Setlow, B., Setlow, P. (1987). Thymine-containing dimers as well as spore photoproducts are found in ultraviolet-irradiated *Bacillus subtilis* spores that lack small, acid-soluble proteins. *Proc. Natl. Acad. Sci. USA* **84,** 421-423.

Setlow, P. (1988). Small acid-soluble, spore proteins of *Bacillus* species: structure, synthesis, genetics, function and degradation. *Annu. Rev. Microbiol*. **42,** 319-338.

Setlow, P. (1993). Spore structural proteins. *In*: Hoch, J.A., Losick, R., Sonenshein, A.L. (Eds.) *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology, and molecular genetics. American Society for Microbiology, Washington, DC, pp. 801-809.

Shazand, K., Frandsen, N., Stragier, P. (1995). Cell-type specificity during development in *Bacillus subtilis*: the molecular and morphological requirements for σ^E activation. *EMBO J.* **14,** 1439-1445.

Spizizen, J. (1958). Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc. Natl. Acad. Sci. USA* **44,** 1072-1078.

Steinmetz, M., Richter, R. (1994). Plasmids designed to alter the antibiotic resistance expressed by insertion mutations in *Bacillus subtilis*, through *in vivo* recombination. *Gene* **142,** 79-83.

Stephens, M.A., Lang, N., Sandman, K.L., Losick, R. (1984). A promoter whose utilization is temporally regulated during sporulation in *Bacillus subtilis*. *J. Mol. Biol.* **176,** 333-348.

Sterlini, J.M., Mandelstam, J. (1969). Commitment to sporulation in *Bacillus subtilis* and its relationship to the development of actinomycin resistance. *Biochem. J*. **113,** 29-37.

Stragier, P., Bonamy, C., Karmazyn-Campbelli, C. (1988). Processing of a sporulation sigma factor in *Bacillus subtilis*: how morphological structure could control gene expression. *Cell* **52,** 697-704.

Sun, D., Stragier, P., Setlow, P. (1989). Identification of a new σ-factor involved in compartmentalized gene expression during sporulation of *Bacillus subtilis*. *Genes Dev*. **3,** 141-149.

Sun, D., Fajardo-Cavazos, P., Sussman, M.D., Tovar-Rojo, F., Cabrera-Martinez, R.-M., Setlow, P. (1991). Effect of chromosome location of *Bacillus subtilis* forespore genes on their *spo* gene dependence and transcription by $E\sigma^F$: identification of features of good $E\sigma^F$-dependent promoters. *J. Bacteriol.* **137,** 7867-7874.

Youngman, P., Perkins, J., Losick, R. (1984). Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in *Bacillus subtilis* or expression of the transposon-borne *erm* gene. *Plasmid* **12,** 1-9.

Chapter IX

Bacillus subtilis contains four closely related Type I signal peptidases with overlapping substrate specificities; constitutive and temporally controlled expression of different *sip* genes

IX.1. Summary

Most biological membranes contain one, or two type I signal peptidases for the removal of signal peptides from secretory precursor proteins. In this respect, the Gram-positive bacterium *Bacillus subtilis* seems to be exceptional, because it contains at least four chromosomally-encoded type I signal peptidases, denoted SipS, SipT, SipU and SipV. Here, we report the identification of the *sipT* and *sipV* genes, and the functional characterization of SipT, SipU and SipV. The four signal peptidases have similar substrate specificities, as they can all process the same ß-lactamase precursor. Nevertheless, they seem to prefer different pre-proteins, as indicated by studies on the processing of the pre-α-amylase of *B. amyloliquefaciens* in strains lacking either SipS, SipT, SipU or SipV. The *sipU* and *sipV* genes are constitutively transcribed at a low level, suggesting that they are required for processing of (pre-)proteins secreted during all growth phases. In contrast, the transcription of *sipS* and *sipT* is temporally controlled, in concert with the expression of the genes for most secretory proteins, which suggests that SipS and SipT serve to increase the secretory capacity of *B. subtilis*. Taken together, our findings suggest that SipS, SipT, SipU, and SipV serve different functions during the exponential and post-exponential growth phase of *B. subtilis*.

IX.2. Introduction

Bacterial proteins that are exported from the cytoplasm through the general pathway for protein secretion are synthesized as precursors with an amino-terminal signal peptide. The signal peptide is required for the targeting of precursor proteins to the cytoplasmic membrane, and for the initiation of their translocation across this membrane. During, or shortly after the translocation process, most signal peptides are removed by type I signal peptidases (SPases), which is a prerequisite for the release of secretory proteins from the extracytoplasmic side of the membrane (1-4).

Homologous type I SPases have been identified in Gram-positive and Gram-negative

bacteria, the inner membrane of yeast mitochondria, and the endoplasmic reticular (ER) membranes of yeast and higher eukaryotes (5-8). Despite the fact that considerable similarities can be observed between the known type I SPases when individual amino acid sequences are compared, only few residues are strictly conserved in all known enzymes. These include serine and lysine residues, which are essential for enzymatic activity, possibly by forming a catalytic dyad (9-12).

Based on computer-assisted analyses (5 and 6; our unpublished results), and studies on the membrane topology of type I SPases of *Escherichia coli*, yeast mitochondria and the canine ER (7, 13-15), it is predicted that the active sites of the known type I SPases are located either in the periplasm of Gram-negative bacteria, the cell wall of Gram-positive bacteria, the mitochondrial intermembrane space, or the lumen of the ER. Nevertheless, based on topological criteria, these enzymes can be divided into four distinct groups. SPases of the first and, apparently, largest group are type II membrane proteins with one amino-terminal membrane anchor. This group includes all known type I SPases from Gram-positive bacteria (6, 16-19), cyanobacteria (20; GenBank accessions D90899 and D90904); and the putative catalytic subunits of the ER type I SPases (8). Furthermore, at least one type I SPase from a Gram-negative bacterium (*ie.* SipS of *Bradyrhizobium japonicum*; Ref. 21), and one from mitochondria (*ie.* Imp1p; Ref. 22) belong to this group. Type I SPases of the second group, which have two amino-terminal membrane anchors, have been identified exclusively in Gram-negative bacteria (23-25). The type I SPase of *Haemophilus influenzae* (26) is the only known representative of the third group, having three putative amino-terminal membrane anchors. Finally, SPases of the fourth group seem to have one amino-terminal and one carboxyl-terminal membrane anchor (our unpublished results). Enzymes of the latter group have been identified in the Gram-negative bacterium *Rhodobacter capsulatus* (27) and yeast mitochondria (*ie.* Imp2p; Ref. 28).

The SPase I of *E. coli*, also known as leader peptidase (Lep), is essential for cell viability (29), and SPase limitation results in the accumulation of precursors of exported proteins (30, 31). Similarly, the type I SPases SpsB from *S. aureus* (18), and Sec11p of the yeast ER membrane (32) are essential enzymes for cell viability. In contrast, the type I SPase SipS of *Bacillus subtilis* (SipS [Bsu]) is not essential for cell viability, and mutant *B. subtilis* strains with a disrupted *sipS* gene are still able to process secretory pre-proteins (6, 33). This suggested the presence of at least one additional type I SPase in *B. subtilis*. Support for the latter hypothesis was first obtained through the identification of two genes for homologous, but non-identical SPases, denoted SipS (Bam) (17) and SipT (Bam) (also known as SipS2), in the closely related bacterium *B. amyloliquefaciens* (16). A few months later, genome sequencing analyses revealed the presence of two open reading frames, *ycsB* (34) and *yhjF* (M.A.N and S.B., unpublished results) from *B. subtilis*, the deduced amino acid sequences of which showed a high degree of similarity to that of SipS (Bsu). By analogy to other SPase-encoding genes of bacilli, we renamed the latter open reading frames *sipU* and *sipV*, respectively, although SPase activity of the corresponding proteins had not been demonstrated.

SipS (Bsu) and SipS (Bam) appeared to be equivalent enzymes in *B. subtilis* and *B. amyloliquefaciens*: first, it was shown that the amino acid sequences of both enzymes are highly similar (91% identical residues and conservative replacements; Ref. 17); and, second, in both organisms the corresponding *sipS* genes were mapped immediately upstream of the *rib* operons for riboflavin biosynthesis (35; our unpublished observations). In contrast, SipT from *B. amyloliquefaciens* showed a much lower degree of sequence similarity to SipU and SipV of *B. subtilis* (65% and 44% identical residues and conservative replacements, respectively), and the corresponding genes were mapped at different regions of the respective chromosomes (16, 34; M.A.N and S.B., unpublished results). The latter findings suggested that SipT from *B. amyloliquefaciens* is not the equivalent of SipU or SipV from *B. subtilis* and, consequently, it was conceivable that these organisms contain at least four chromosomal *sip* genes for type I SPases. The present studies were aimed at the verification of this hypothesis, and the functional characterization of the type I SPases of *B. subtilis*. We show that *B. subtilis* contains four closely related type I SPases which have similar, but non-identical substrate specificities. In addition, we show that the transcription of the *sipT* gene is temporally controlled, whereas the *sipU* and *sipV* genes are constitutively transcribed at a low level.

IX.3. Experimental procedures

Plasmids, bacterial strains and media

Table 1 lists the plasmids and bacterial strains used.

Table IX.1. Plasmids and bacterial strains

Plasmids	Relevant properties	Ref.
pGDL41	Encodes pre(A13i)-ß-lactamase and SipS (Bsu); replicatesin *E. coli* and *B. subtilis;* 8.1 kb; Ap^r; Em^s; Km^r	(6)
pGDL48	Lacks the *sipS* (Bsu) gene and contains a multiple cloning site; otherwise identical to pGDL41; 7.5 kb; Ap^r; Km^r	(17)
pGDL100	pGDL48 derivative carrying the *sipT* (Bsu) gene; 8.3 kb	This paper
pGDL110	pGDL48 derivative carrying the *sipT* (Bam) gene; 8.3 kb	This paper
pGDL121	pGDL48 derivative carrying the *sipU* (Bsu) gene; 8.6 kb	This paper
pGDL131	pGDL48 derivative carrying the *sipV* (Bsu) gene; 8.3 kb	This paper
pGDL132	pGDL48 derivative carrying the *sipV-myc* (Bsu) gene; 8.3 kb	This paper
pHT100C	pUC18 derivative for the disruption of *sipT* (Bsu)*;* 4.5 kb; Ap^r; Cm^r	This paper
pORI280	Integration vector carrying the *lacZ* gene of *E. coli*; unable to replicate in *B. subtilis*; 4.4 kb; Em^r	(36)
pINT34d	pORI280 derivative for the deletion of a 197-bp *Eco*RI fragment with the 5' end of *sipU* (Bsu); contains upstream sequences of *sipU* (0.3 kb)and the 3' end of *sipU*, but lacks the 197-bp *Eco*RI fragment; 5.0 kb	This paper
pV50E	pUC18 derivative for the disruption of *sipV* (Bsu)*;* 4.4 kb; Ap^r; Em^r	This paper
pKTH10	Encodes the a-amylase (AmyQ) of *B. amyloliquefaciens*, 6.8 kb; Km^r	(37)
pLGW200	Integration vector for *B. subtilis*; contains a promoterless *lacZ* gene fused to the ribosome-binding site of the *spoVG* gene; 6.8 kb. Cm^r	(38)
pGDE22	pLGW200 derivative with a transcriptional *sipS-lacZ* fusion; 7.5 kb	(33)
pLGT207	pLGW200 derivative with a transcriptional *sipT-lacZ* fusion; 7.6 kb	This paper
pLGU202	pLGW200 derivative with a transcriptional *sipU-lacZ* fusion; 7.8 kb	This paper
pLGV201	pLGW200 derivative with a transcriptional *sipV-lacZ* fusion; 7.7 kb	This paper

Strains	Genotype	Ref.
E. coli		
MC1061	F⁻; *araD*139; D(*ara-leu*)7696; D(*lac*)X74; *galU*; *galK*; *hsdR2*; *mcrA*; *mcrB1*; *rspL*	(39)
B. subtilis		
168	*trpC2*	provided by C. Anagnostopoulos
8G5	*trpC2; tyr; his; nic; ura; rib; met; ade*	(40)
8G5 *sipS*	*trpC2; tyr; his; nic; ura; met; ade; sipS*	(33)
8G5*sipT*-Cm	*trpC2; tyr; his; nic; ura; rib; met; ade; sipT; Cm*^{r}	This paper
8G5 *sipU*	*trpC2; tyr; his; nic; ura; rib; met; ade; sipU*	This paper
8G5 *sipV-E*m	*trpC2; tyr; his; nic; ura; rib; met; ade; sipV; Em*^{r}	This paper
8G5::pGDE22	8G5 carrying pGDE22 (*sipS-lacZ*) in the chromosome; Cm^r	(33)
8G5::pLGT207	8G5 carrying pLGT207 (*sipT-lacZ*) in the chromosome; Cm^r	This paper
8G5::pLGU202	8G5 carrying pLGU202 (*sipU-lacZ*) in the chromosome; Cm^r	This paper
8G5::pLGV201	8G5 carrying pLGV201 (*sipV-lacZ*) in the chromosome; Cm^r	This paper

TY medium (tryptone/yeast extract) contained Bacto tryptone (1%), Bacto yeast extract (0.5%) and NaCl (1%). Minimal medium for *B. subtilis* was prepared as described in Ref. 41, and supplemented with glucose (0.5%), casamino acids (0.02%), tryptophan (20 μg/ml), histidine (20 μg/ml), methionine (20 μg/ml), tyrosine (20 μg/ml), adenine (20 μg/ml), uracil (20 μg/ml), nicotinic acid (0.4 μg/ml), riboflavin (0.4 μg/ml), and Fe-ammoniumcitrate (1.1 μg/ml). M9 media 1 and 2, used for the labeling of *E. coli* proteins with [^{35}S]-methionine (Amersham International plc, Little Chalfont, UK), were prepared as described in Ref. 42. S7 media 1 and 3, used for labeling of *B. subtilis* proteins, were prepared as described in Ref. 43. If required, media for *E. coli* were supplemented with ampicillin (50 μg/ml), erythromycin (100 μg/ml), or kanamycin (40 μg/ml); media for *B. subtilis* were supplemented with chloramphenicol (5 μg/ml), erythromycin (2 μg/ml), or kanamycin (10 μg/ml).

DNA techniques

Procedures for DNA purification, restriction, ligation, agarose gel electrophoresis, and transformation of *E. coli* were carried out as described in Ref. 44. Enzymes were from Boehringer (Mannheim, Germany). Competent *B. subtilis* cells were transformed as described in Ref. 40. The sequences of DNA fragments, including PCR-amplified fragments, were analyzed by the dideoxy-chain termination method (45), using the T7 Sequencing Kit (Pharmacia, Uppsala, Sweden). [^{35}S]-dATP (8 μCi/μl; > 1000 Ci/mmole) from Amersham International. DNA and protein sequences were analyzed using version 6.7 of the PCGene Analysis Program (Intelligenetics Inc., Mountain View, CA). The BLASTP algorithm (46) was used for protein comparisons in GenBank. Correct integration of linearized DNA fragments, or plasmids in the chromosome of *B. subtilis* was verified by Southern hybridization. PCR under stringent conditions for the annealing of primers to template DNA was carried out with Vent DNA polymerase (New England Biolabs, Beverly, USA) as described in Ref. 11. When low-stringency conditions were required, the annealing temperature was lowered to 42°C, and the Supertaq DNA polymerase (Sphaero-Q, Leiden, the Netherlands) was used.

Pulse-chase protein labeling, immunoprecipitation, SDS-PAGE and fluorography

Pulse-chase labeling of *E. coli* and *B. subtilis*, immunoprecipitation, SDS-PAGE and fluorography were performed as described in Refs. 42 and 43. [^{14}C]-methylated molecular weight markers were from Amersham International. Relative amounts of precursor and mature forms of secreted proteins were estimated by film scanning with an LKB ultroscan XL laser densitometer (LKB, Bromma, Sweden).

ß-galactosidase activity assay

Overnight cultures were diluted 100-fold in fresh medium and samples were taken hourly for optical density (OD) readings at 600 nm and ß-galactosidase activity determinations. The assay and the calculation of ß-galactosidase units (expressed as units per OD600) were carried out as described in Ref. 47.

Western blot analysis

Western blotting was performed as in Ref. 48. After separation by SDS-PAGE, proteins were transferred to Immobilon-PVDF membranes (Millipore Corporation, Bedford, MA, USA). To assay the presence of the precursor and mature forms of *B. amyloliquefaciens* α-amylase in *B. subtilis*, cells were separated from the growth medium by centrifugation (5 min, 12.000 rpm, room temperature), and samples for SDS-PAGE were prepared as described in Ref. 42. The *B. amyloliquefaciens* α-amylase was visualized with specific antibodies and horseradish peroxidase-anti-rabbit-IgG conjugates (Amersham International). To monitor the presence of the SipV-Myc fusion protein in *E. coli*, samples for SDS-PAGE were prepared as in Ref. 11. SipV-Myc was visualized with specific monoclonal antibodies and horseradish peroxidase-anti-mouse-IgG conjugates (Amersham International).

IX.4. Results

Identification of the *sipT* gene of *B. subtilis*

To determine whether *B. subtilis* contains a *sipT* (Bam)-like gene, Southern hybridization experiments were performed. A 3.2-kb *Hin*dIII fragment of *B. subtilis* chromosomal DNA was found to hybridize weakly with the *sipT* (Bam) gene, but not with the *sipS* (Bsu), *sipU* (Bsu), or *sipV* (Bsu) genes, suggesting that *B. subtilis* contains a *sipT* (Bam)-like gene (data not shown). To identify the latter gene, a PCR was performed with the primers lba2-1 and lba2-2 (Fig. IX.1), which correspond to sequences within the *sipT* (Bam) gene. Using chromosomal DNA of *B. subtilis* 8G5 *sipS* as a template, a 300-bp fragment was amplified. Sequence analysis of this fragment revealed the presence of an open reading frame, the deduced amino acid sequence of which showed a high degree of similarity to SipT (Bam) and, to a lesser extent, to SipS (Bsu), SipU (Bsu), and SipV (Bsu) (data not shown). The latter findings indicated that the 300-bp fragment was an internal fragment of a fourth *sip* gene of *B.*

subtilis, possibly *sipT*. The entire *sipT* (Bam)-like gene of *B. subtilis* was cloned in three successive PCR steps (schematically shown in Fig. IX.1, b-d). Sequence analysis showed that the upstream sequences of this gene contain the 3' end of the *B. subtilis fruA* gene for the enzyme II of the fructose-specific phosphoenol pyruvate phosphotransferase system (Fig. IX.1), which has been mapped at 126 degrees of the *B. subtilis* chromosome (49). The latter finding showed that the fourth *sip* gene of *B. subtilis* is indeed the *sipT* (Bsu) gene (GenBank accession U45883), because the *sipT* (Bam) gene is also preceded by *fruA* on the chromosome of *B. amyloliquefaciens* (16).

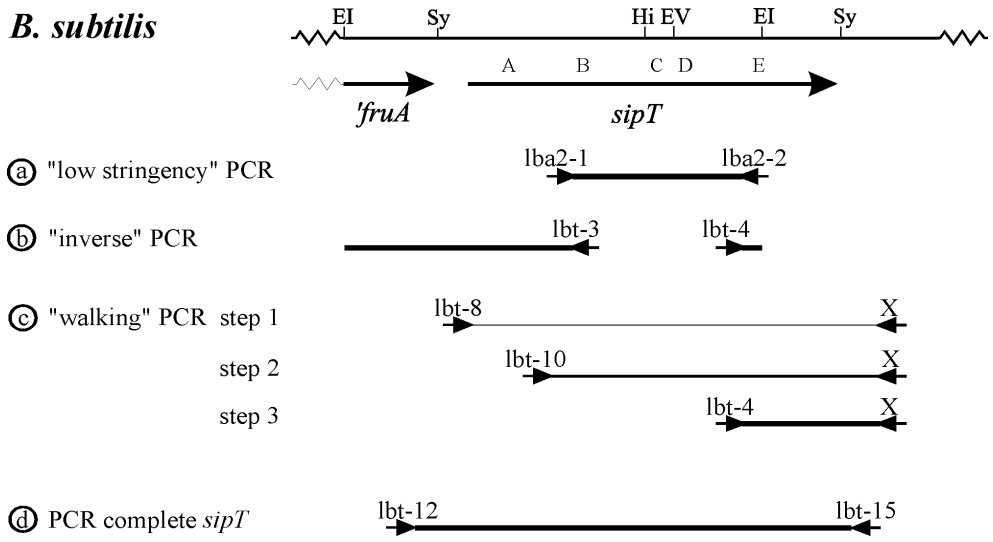


Fig. IX.1. Identification of the *B. subtilis sipT* gene. To clone the *sipT* (Bsu) gene, four subsequent PCR steps (a-d) were performed with genomic DNA of *B. subtilis* as a template. The nucleotide sequences (5'-3') of primers used for PCR are indicated below; nucleotides identical to genomic template DNA are printed in bold, and restriction sites used for cloning are underlined. After each PCR (a-d), amplified fragments were sub-cloned and sequenced. In step a, an internal, 300-bp fragment of *sipT* (Bsu) was amplified with primers lba2-1 (**TGAACCGTACTTAGTGG**) and lba2-2 (**GATTGTCGCCCATGACG**) under conditions of low stringency. In step b, a 600-bp fragment containing the 5' end of *sipT* (Bsu) was amplified by "inverse PCR" with primers lbt-3 (TTGAA**TTC**ACAAACAGCC**TTTCTCCG**) and lbt-4 (GAGAATTC**GGACCGGTTAAGGTTCCG**), using a self-ligated, circular, 0.8-kb genomic *Eco*RI fragment as the template. In step c, a 250-bp fragment containing the 3' end of *sipT* (Bsu) was amplified from the *B. subtilis* genome by "walking PCR" using the primers lbt-4 (see above), lbt-8 (GATAGTCGAC**AAAGAAGAGAAACAACTG**), and lbt-10 (GGAAGTCGAC**ATACGTACCTGGAATGG**), in combination with a set of primers (X) with the following sequences: GGAAGATCTGAATTCATAAAGGGAAGATG; GCGAATTCTTTTATCAGCGTTCTGGCT; TTTGAATTCTACTTACTGTCACTCGTT; GATCGAATTCGATGGCGCTACTCTGGG; GATCGAATTCATAAAGAACTAAACCTCGGTG. In step c, a first PCR was performed under conditions of low stringency with primers lbt-8 and X. This resulted in the amplification of a wide range of different fragments, which were used as template DNA in a second PCR under stringent conditions with primers lbt-10 and X. The products of the second PCR were used as template for a third PCR with primers lbt-4 and X. Finally, in step d, an 862-bp DNA fragment containing the complete *sipT* gene of *B. subtilis* 168 was amplified with primers lbt-12 (GATGGTCGAC**CTTTTAAGTATCGTGATCGG**) and lbt-15 (CACGGTACC**ATGCATTGCATTGGTCGC**). The sequence of *sipT* (Bsu) was determined from three independent isolates (GenBank accession U45883). Regions of *sipT* specifiying the conserved domains (A-E), and relevant restriction sites are shown: EI, *Eco*RI; EV, *Eco*RV; Hi, *Hind*II; Sy, *Sty*I.

Like other *sip* genes from bacilli (6, 16, 17, 34), the *sipT* (Bsu) gene (582 nucleotides) is preceded by a potential ribosome binding site (GGAGG); it has a TTG startcodon, and lacks upstream sequences with obvious similarity to the major classes of *B. subtilis* promoters. The deduced amino acid sequence of the SipT (Bsu) protein (193 residues), shows a high degree of similarity to that of the known type I SPases from *B. subtilis* and related bacilli (Fig. IX.2). Like the other *Bacillus* enzymes, SipT (Bsu) belongs to the sub-group of SPases with one amino-terminal membrane anchor. Furthermore, SipT (Bsu) contains the five conserved domains (A-E; Ref. 6) that are present in all known type I SPases, and include the conserved serine (domain B) and lysine (domain D) residues implicated in catalysis (Fig. IX.2; Ref. 11).

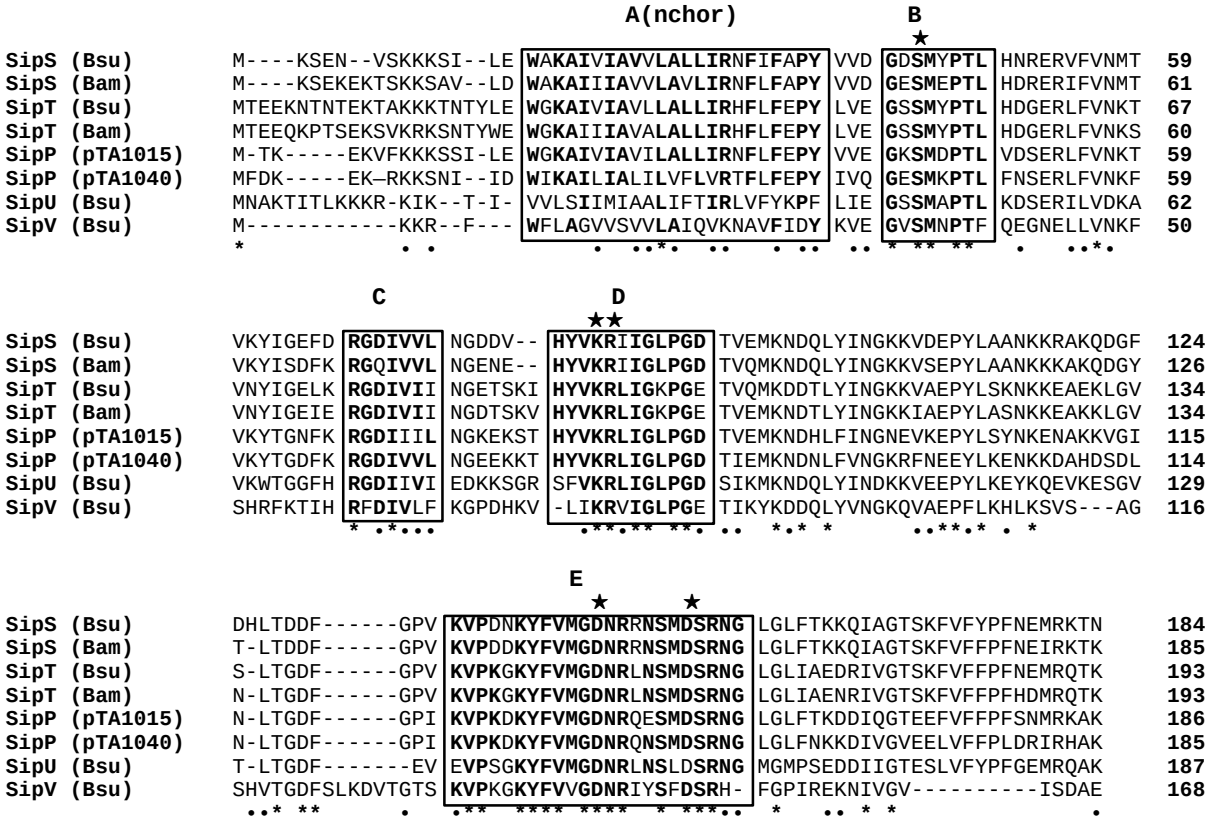


Fig. IX.2. Type I SPases of *B. subtilis* and *B. amyloliquefaciens*. Comparison of the deduced amino acid sequences of type I SPases from *B. subtilis* and *B. amyloliquefaciens*. The comparison includes SipS (Bsu) (6); SipS (Bam) (17); SipT (Bsu); SipT (Bam) (16); the plasmid-encoded type I SPases SipP (pTA1015) and SipP (pTA1040) of *B. subtilis* (*natto*) (17); SipU(Bsu) (34); and SipV (Bsu). Identical amino acids [*], or conservative replacements [•], are marked. Putative transmembrane segments, indicated with A(nchor), were predicted as described in Ref. 57. The conserved domains B-E, which are present in all known type I SPases of prokaryotic and eukaryotic organisms (6), are indicated. Conserved residues which are critical for the activity and/or stability of SipS (Bsu) (11) are marked (★).

The *sipT, sipU,* and *sipV* genes specify functional type I SPases

The hybrid precursor pre(A13i)-ß-lactamase, specified by plasmid pGDL48, is processed by SipS (Bsu) (6), and by plasmid-encoded type I SPases (SipP) of *B. subtilis* (17), but not by Lep of *E. coli*. To test the functionality of SipT (Bsu), SipT (Bam), SipU (Bsu), and SipV

(Bsu), the corresponding genes were amplified by PCR and cloned in pGDL48, resulting in pGDL100, pGDL110, pGDL121, and pGDL131, respectively. *E. coli* MC1061 was transformed with these plasmids and processing of pre(A13i)-ß-lactamase was analyzed by pulse-chase labeling. In addition, *E. coli* MC1061 was transformed with pGDL41, carrying the *sipS* (Bsu) gene (positive control), or pGDL48 (no *sip* gene; negative control). The results showed that after a chase of 10 min significant amounts of pre(A13i)-ß-lactamase were processed to the mature form in *E. coli* strains transformed with pGDL41, pGDL100, pGDL110, or pGDL121, but not in *E. coli* strains transformed with pGDL48 (Fig. IX.3*A*) or (pGDL131) (result not shown). These observations demonstrate that the *sipT* (Bsu), *sipT* (Bam), and *sipU* (Bsu) genes specify functional SPases that can cleave pre(A13i)-ß-lactamase when produced in *E. coli*.

Pre(A13i)-ß-lactamase was processed more efficiently in *E. coli* cells containing *sipT* (Bsu), or *sipT* (Bam) than in cells containing *sipS* (Bsu), or *sipU* (Bsu). In the *sipT* (Bsu or Bam) containing cells, approximately 85% of the (A13i)-ß-lactamase was mature after a chase of ten minutes (Fig. IX.3*A*). In contrast, under the same conditions, about 40% of the (A13i)-ß-lactamase was mature in *E. coli* cells containing *sipS* (Bsu) or *sipU* (Bsu) (Fig. IX.3*A*). These differences in processing activity may either reflect differences in the production of SipS, SipT and SipU, or differences in their ability to cleave pre(A13i)-ß-lactamase.

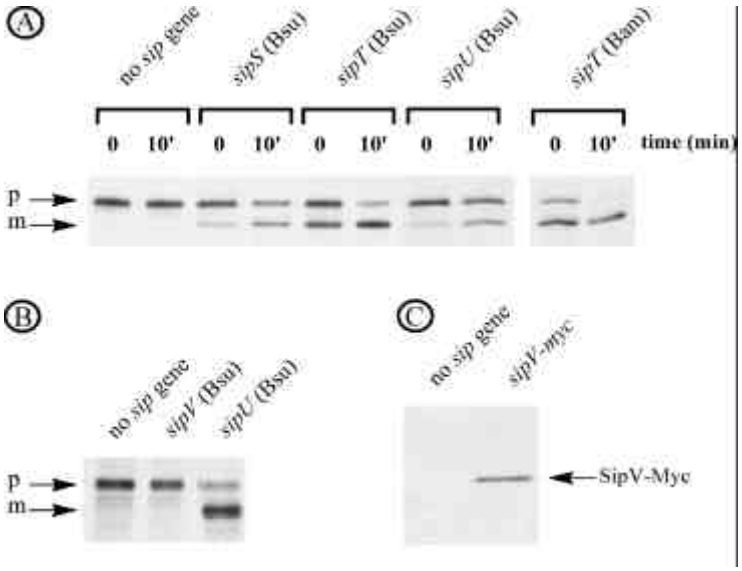


Fig. IX.3. Processing of pre(A13i)-ß-lactamase in *E. coli*. *A,* Processing of pre(A13i)-ß-lactamase in *E. coli* MC1061, transformed with pGDL48 (no *sip* gene), pGDL41 (*sipS* [Bsu]), pGDL100 (*sipT* [Bsu]), pGDL121 (*sipU* [Bsu]), or pGDL110 (*sipT* [Bam]), was analyzed by pulse-chase labeling at 37°C and subsequent immunoprecipitation, SDS-PAGE and fluorography. Cells were labeled with [^{35}S]-methionine for 1 min prior to chase with excess of non-radioactive methionine. Samples were withdrawn at the time of chase (t=0), and 10 min after the chase (t=10). Variations in the amounts of (A13i)-ß-lactamase precipitated from different strains relate only to variability in the incorporation of label into cells of different cultures and not to specific effects of the different SPases. *p*, precursor; *m*, mature. ***B,*** Processing of pre(A13i)-ß-lactamase in *E. coli* MC1061, transformed with pGDL48 (no *sip* gene), pGDL131 (*sipV* [Bsu]), or pGDL121 (*sipU* [Bsu]) was analyzed as in *A*. Samples were withdrawn 60 min after the chase. *p*, precursor; *m*, mature. ***C,*** Production of SipV-Myc in *E. coli* MC1061. The oligonucleotide lbv-7 (TTGAATTCCTTAGTT-CAAATCTTCCTCACTGATCAATTTCTG**TTCGGCATCAGAAATCACACCG**), specifying the human c-Myc epitope was fused in frame to the3' end of *sipV*. Cells transformed with the plasmids pGDL48 (no *sip* gene) or pGDL132 (*sipV-myc*) were grown overnight in TY medium and analyzed by SDS-PAGE and Western blotting. The position of SipV-Myc is indicated.

Because, after a chase of 10 min, no mature (A13i)-ß-lactamase could be detected in *E. coli* (pGDL131) cells containing the cloned *sipV* (Bsu) gene, pulse-chase labeling experiments were performed in which the chase was extended to 60 min. Under these conditions, approximately 90% of the (A13i)-ß-lactamase was processed to the mature form in cells containing SipU (positive control) (Fig. IX.3*B*). In contrast, no mature (A13i)-ß-lactamase could be detected in *E. coli* (pGDL48; negative control), and *E. coli* (pGDL131; *sipV* [Bsu]) (Fig. IX.3*B*). To verify that the presence of pGDL131 resulted in the production of SipV in *E. coli*, an oligonucleotide specifying the human c-myc epitope (EQKLISEEDLN, Ref. 50) was fused in frame to the 3' end of *sipV*, resulting in pGDL132. As shown by Western blotting, the SipV-Myc protein was produced in *E. coli* cells containing pGDL132 (Fig. IX.3*C*). Pulse-chase labeling experiments revealed that pre-(A13i)-ß-lactamase was not processed in *E. coli* (pGDL132), like in *E. coli* (pGDL131) (data not shown). The latter findings indicate that SipV is produced in *E. coli* cells containing pGDL131, and that SipV is unable to cleave pre(A13i)-ß-lactamase in *E.coli*.

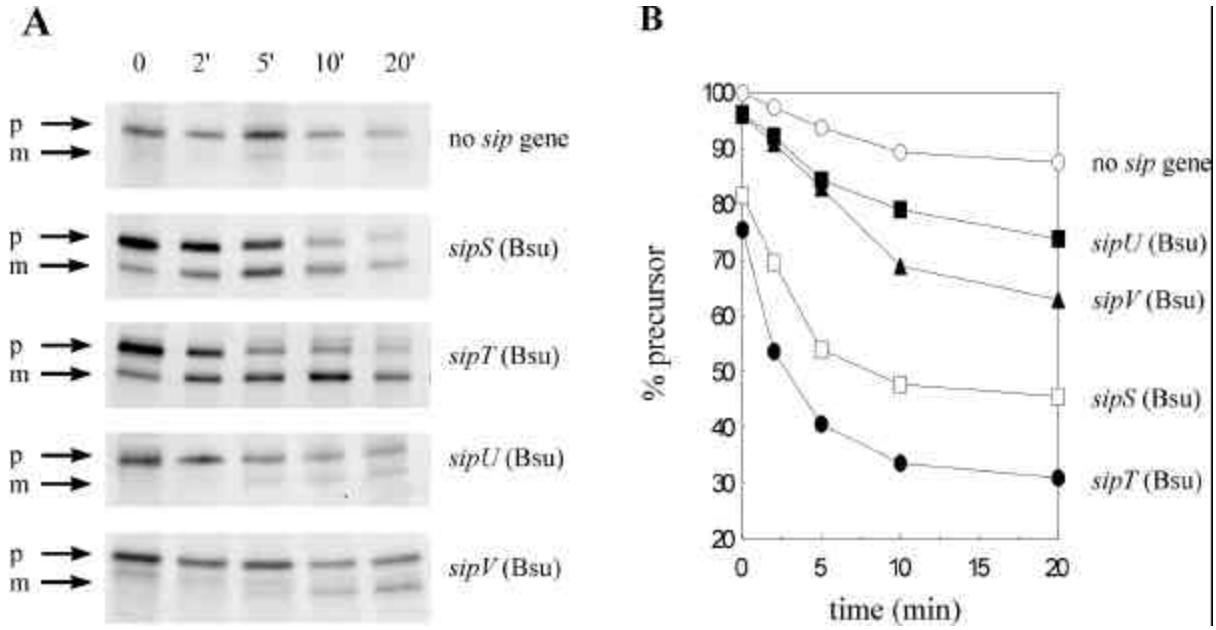


Fig. IX.4. Processing of pre(A13i)-ß-lactamase in *B. subtilis*. *A*, Processing of pre(A13i)-ß-lactamase in *B. subtilis* 8G5 *sipS* harbouring pGDL48 (no *sip* gene), pGDL41 (*sipS* [Bsu]), pGDL100 (*sipT* [Bsu]), pGDL121 (*sipU* [Bsu]), or pGDL131 (*sipV* [Bsu]) was analyzed by pulse-chase labeling at 37°C and subsequent immunoprecipitation, SDS-PAGE and fluorography. Cells were labeled with [^{35}S]-methionine for 1 min prior to chase with excess of non-radioactive methionine. Samples were withdrawn at the times indicated. The loss of label in the time courses is due both to removal of the signal peptide and degradation of the mature ß-lactamase in the medium (58). ***B***, The kinetics of processing are plotted as the percentage of the total (A13i)-ß-lactamase (precursor + mature) present in the precursor form at the time of sampling. Relative amounts of the precursor and mature forms of pre(A13i)-ß-lactamase were determined by scanning of autoradiographs. O, pGDL48 (no *sip* gene); □, pGDL41 (*sipS* [Bsu]); , pGDL100 (*sipT* [Bsu]); , pGDL121 (*sipU* [Bsu]); , pGDL131 (*sipV* [Bsu]).

To analyze pre(A13i)-ß-lactamase processing by SipS, SipT, SipU, and SipV in *B.*

subtilis, the *B. subtilis* strain 8G5 *sipS* was transformed with pGDL41 (*sipS* [Bsu]), pGDL100 (*sipT* [Bsu]), pGDL121 (*sipU* [Bsu]), pGDL131 (*sipV* [Bsu]), and pGDL48 (no *sip* gene), respectively. As reported previously (33), pre(A13i)-ß-lactamase was processed at a low rate in *B. subtilis* 8G5 *sipS* (pGDL48). In contrast, pre(A13i)-ß-lactamase was processed at significantly increased rates in *B. subtilis* 8G5 *sipS* strains with plasmids carrying *sipS* (Bsu) or *sipT* (Bsu). As compared to the latter two strains, the presence of a plasmid with the *sipU* (Bsu) gene (*ie*. pGDL121) resulted in a less drastic increase in the rate of pre(A13i)-ß-lactamase processing (Fig. IX.4, *A* and *B*). Surprisingly, the presence of a plasmid with the *sipV* (Bsu) gene (*ie*. pGDL131) also resulted in an increased rate of pre(A13i)-ß-lactamase processing (Fig. IX.4, *A* and *B*), suggesting that SipV can process this precursor in *B. subtilis* but, as described in the foregoing section, not in *E. coli*.

SipT, SipU, or SipV are not essential for cell viability

It was previously shown that SipS is not essential for viability of *B. subtilis*, and cells lacking the *sipS* gene were still able to process secretory pre-proteins (33). To investigate whether SipT is essential for viability of *B. subtilis*, the following strategy was used: first, a plasmid-encoded copy of the corresponding gene was disrupted with a chloramphenicol resistance (Cm^r) marker. The resulting plasmid pHT100C, which is unable to replicate in *B. subtilis*, was linearized and, subsequently, used to transform competent *B. subtilis* 8G5 cells. As verified by Southern hybridization (data not shown), all chloramphenicol-resistant transformants (denoted *B. subtilis* 8G5 *sipT*-Cm) contained the disrupted *sipT* gene (schematically presented in Fig. IX.5*A*), showing that SipT is not required for cell viability.

Similarly, it was shown that SipU is not essential for cell viability by deleting a 197-bp *Eco*RI fragment from the chromosome of *B. subtilis* (schematically shown in Fig. IX.5*A*). The latter fragment contains the first 170 bp of the *sipU* gene specifying the conserved domains A (*ie*. the membrane anchor) and B (containing the putative catalytic serine residue; Fig. IX.2). To this purpose, we used plasmid pINT34d, which consists of the chromosomal integration plasmid pORI280 (36) carrying a mutant copy of the *sipU* locus that lacks the 197-bp *Eco*RI fragment. Upon the Campbell-type integration of pINT34d into the *sipU* locus of the *B. subtilis* chromosome, and the subsequent selection of cells that had lost this plasmid from the chromosome, it was shown by PCR and Southern blotting that about 10% of the cells lacking pINT34d also lacked the 197-bp *Eco*RI fragment. This finding showed that SipU is not essential for cell viability. The resulting mutant strain was denoted *B. subtilis* 8G5 *sipU*.

To disrupt the chromosomal *sipV* gene, we first disrupted a plasmid-encoded copy of this gene with an erythromycin resistance (Em^r) marker. The resulting plasmid pV50E, which is unable to replicate in *B. subtilis*, was linearized and, subsequently, used to transform competent *B. subtilis* 8G5 cells. All erythromycin-resistant transformants (denoted *B. subtilis* 8G5 *sipV*-Em) contained the disrupted *sipV* gene (schematically presented in Fig. IX.5*A*), showing that, like SipS, SipT, and SipU, also SipV is not required for cell viability.

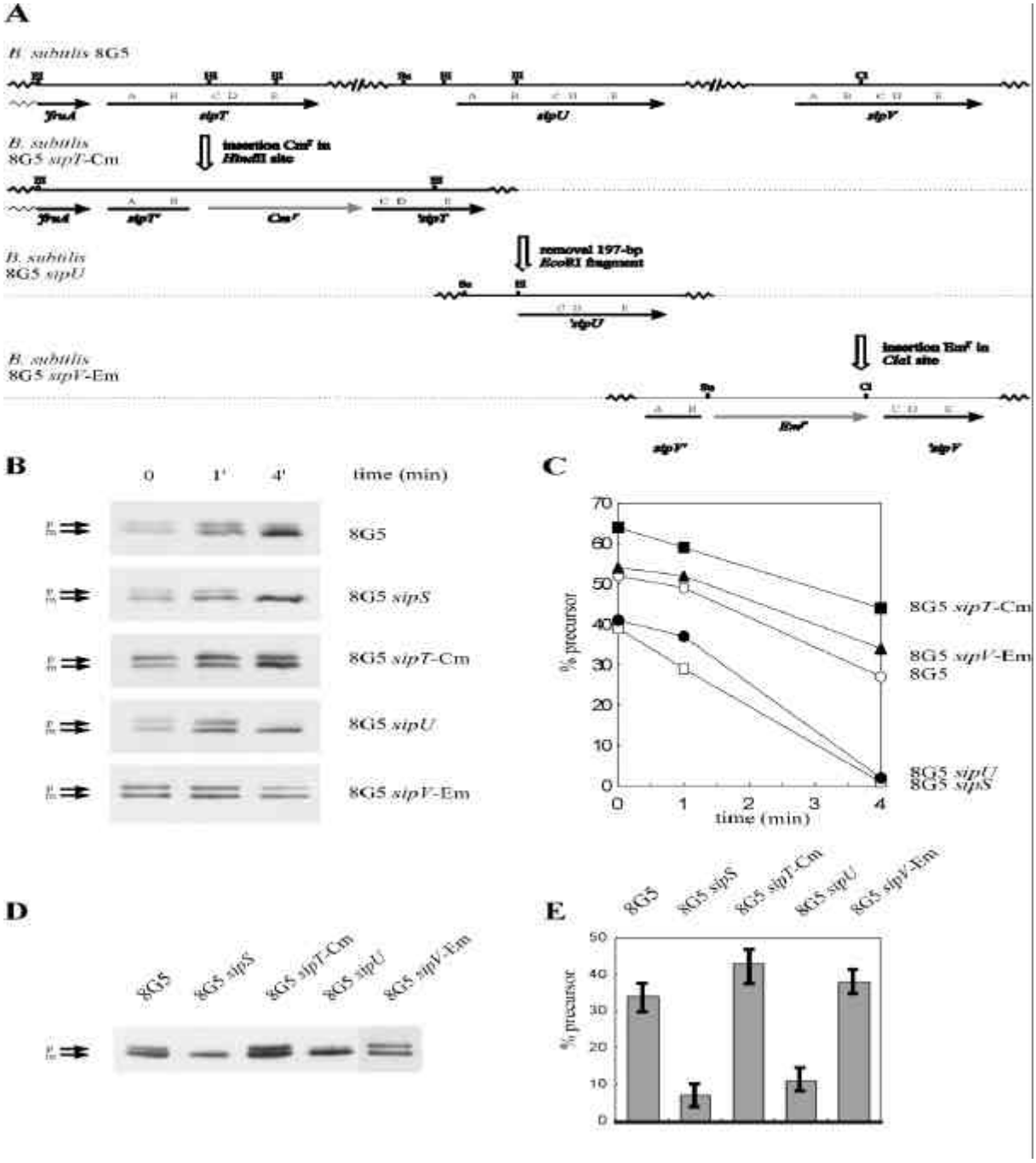


Fig. IX.5. Processing of the *B. amyloliquefaciens* α-amylase precursor in mutant *B. subtilis* strains lacking SipS, SipT, SipU, or SipV. *A*, Schematic presentation of the construction of *B. subtilis* 8G5 *sipT*-Cm, *B. subtilis* 8G5 *sipU* and *B. subtilis* 8G5 *sipV*-Em. The chromosomal *sipT* gene was disrupted with a Cm^r marker by homologous recombination. To this purpose, *B. subtilis* 8G5 was transformed with the linearized plasmid pHT100C, which can not replicate in *B. subtilis*, and which contains a mutant copy of *sipT* with a Cm^r marker in the *Hin*dII site. Part of the *sipU* gene was removed from the chromosome of *B. subtilis* 8G5 using pINT34d, a derivative of the chromosomal integration plasmid pORI280 (36). In addition to an Em^r marker and the *E. coli lacZ* gene, pINT34d carries a mutant copy of *sipU*, which was obtained by the excision of a 197-bp *Eco*RI fragment from a PCR-amplified genomic DNA fragment containing *sipU* and its flanking sequences. Since pINT34d is unable to replicate in *B. subtilis*, transformants (Em^r and blue on plates with X-gal) can only be obtained as a result of a Campbell-type integration into the homologous *sipU* sequences on the chromosome. (*Continued on next page*).

Legend to Fig IX.5, continued
Thus, the chromosome of transformants with pINT34d will contain both an intact, and a truncated copy of *sipU*, separated by sequences of pORI280. Upon growth for about 200 generations in the absence of Em, cells were selected that had spontaneously lost pORI280, together with one of the two copies of *sipU* (Em^s and white on plates with X-gal). Cells lacking the 197-bp *Eco*RI fragment were denoted *B. subtilis* 8G5 *sipU*. The chromosomal *sipV* gene was disrupted with an Em^r marker by homologous recombination. To this purpose, *B. subtilis* 8G5 was transformed with the linearized plasmid pV50E, which can not replicate in *B. subtilis*, and which contains a mutant copy of *sipV* with an Em^r marker in the *Cla*I site. Only restriction sites relevant for the constructions are shown (Cl, *Cla*I; EI, *Eco*RI; Hi, *Hin*dII; Su, *Stu*I). Regions of *sipT*, *sipU* and *sipV* specifying the conserved domains A to E are indicated. ***B* and *C***, Processing of the *B. amyloliquefaciens* α-amylase precursor in *B. subtilis* strains lacking SipS, SipT, SipU, or SipV. Processing of pre-α-amylase in *B. subtilis* 8G5 (pKTH10), *B. subtilis* 8G5 *sipS* (pKTH10), *B. subtilis* 8G5 *sipT*-Cm (pKTH10), *B. subtilis* 8G5 *sipU* (pKTH10), and *B. subtilis* 8G5 *sipV*-Em (pKTH10), was analyzed by pulse-chase labeling at 37°C and subsequent immunoprecipitation, SDS-PAGE and fluorography. Cells were labeled with [^{35}S]-methionine for 1 min prior to chase with excess non-radioactive methionine. Samples were withdrawn at the times indicated. Since the incorporation of label into α-amylase cannot be stopped instantaneously by the addition of non-radioactive methionine, samples withdrawn at t=0 contain less labeled α-amylase than samples withdrawn at t=1 and t=4 min. p, precursor; m, mature. ***C***, Relative amounts of precursor and mature forms of α-amylase were determined by densitometer scanning of autoradiographs. The kinetics of processing are plotted as in Fig. 4*B*. ○, *B. subtilis* 8G5 (pKTH10); □, *B. subtilis* 8G5 *sipS* (pKTH10); ■, *B. subtilis* 8G5 *sipT*-Cm (pKTH10); ●, *B. subtilis* 8G5 *sipU* (pKTH10); and ▲, *B. subtilis* 8G5 *sipV*- Em (pKTH10). ***D* and *E***, Accumulation of pre-α-amylase in cells of *B. subtilis* 8G5 (pKTH10), *B. subtilis* 8G5 *sipS* (pKTH10), *B. subtilis* 8G5 *sipT*-Cm (pKTH10), *B. subtilis* 8G5 *sipU* (pKTH10), and *B. subtilis* 8G5 *sipV*-Em (pKTH10) was analyzed by SDS-PAGE and Western blotting. Cells were grown overnight in TY medium. p, precursor; m, mature. ***E***, Relative amounts of precursor and mature forms of α-amylase accumulating in cells of *B. subtilis* were determined by densitometer scanning of films. The average values of three individual experiments are shown, and the standard deviation is indicated by error bars. In each of the latter experiments *B. subtilis* 8G5 *sipT*-Cm accumulated more pre-α-amylase than *B. subtilis* 8G5, or *B.subtilis* 8G5 *sipV*-Em.

Neither the disruption of the *sipT* or *sipV* genes, nor the removal of an essential part of the *sipU* gene, had a detectable influence on cell growth, the development of competence for DNA binding and uptake, or sporulation (data not shown).

Processing of α-amylase is reduced in the absence of SipT, and improved in the absence of SipS or SipU

Processing of the precursor of the *B. amyloliquefaciens* α-amylase AmyQ (previously also referred to as AmyE; Refs. 33, and 51) was recently shown to be improved in the absence of SipS, indicating that the production of SipS interferes with pre-AmyQ processing, and that this precursor could be a preferred substrate for other SPases, such as SipT, SipU or SipV (33). To investigate the effects of the absence of SipT, SipU or SipV on the processing of pre-AmyQ, *B. subtilis* 8G5 *sipT*-Cm, *B. subtilis* 8G5 *sipU*, and *B. subtilis* 8G5 *sipV*-Em were transformed with plasmid pKTH10. The latter plasmid contains the *amyQ* gene, and its presence in *B. subtilis* results in the secretion of α-amylase at high levels (± 1.3 g/l; Refs. 37, and 51). First, we performed pulse-chase labeling experiments. The results showed that,

compared to the parental strain 8G5, the rate of pre-AmyQ processing in the mutant lacking SipT was reduced; after a chase of 4 min, about 30% of the labeled AmyQ was in the precursor form in the parental 8G5 strain whereas, under the same conditions, about 50% of the AmyQ was in the precursor form in *B. subtilis* 8G5 *sipT*-Cm (Fig. IX.5, *B* and *C*). In contrast, the rate of pre-AmyQ processing was increased in strains lacking either SipS or SipU; both in *B. subtilis* 8G5 *sipS* and *B. subtilis* 8G5 *sipU* only about 2% of the labeled AmyQ was present in the precursor form after a chase of 4 min. Processing of pre-AmyQ was hardly affected in *B. subtilis* 8G5 *sipV*-Em (Fig. IX.5, *B* and *C*).

To compare the effects of the absence of SipS, SipT, SipU, or SipV on the accumulation of pre-AmyQ, Western blotting experiments were performed with cells of *B. subtilis* 8G5 *sipS* (pKTH10), *B. subtilis* 8G5 *sipT*-Cm (pKTH10), *B. subtilis* 8G5 *sipU* (pKTH10), *B. subtilis* 8G5 *sipV*-Em (pKTH10), and the parental strain *B. subtilis* 8G5 (pKTH10), all grown overnight in TY medium. As previously shown for strains lacking SipS (33), compared to the parental strain 8G5, the absence of SipU resulted in a reduction of about 20% in the accumulation of pre-AmyQ (Fig. IX.5, *D* and *E*). In contrast, cells lacking SipT accumulated more pre-AmyQ (approximately 10%) than the parental strain, whereas the absence of SipV had no clear effect on the accumulation of pre-AmyQ (Fig. IX.5, *D* and *E*). Taken together, our findings indicate that pre-AmyQ is a preferred substrate for SipT, and that the presence of SipS or SipU interferes with efficient processing of this precursor. It is not clear whether pre-AmyQ is a substrate for SipV.

Distinct regulation of *sipT*, *sipU*, and *sipV* gene expression at the transcriptional level

The transcription of the *sipS* (Bsu) gene is temporally regulated, *sipS* promoter activity being highest in the post-exponential growth phase (33). To examine whether this is also the case for the transcription of the *sipT*, *sipU* and *sipV* genes, transcriptional *sipT-lacZ*, *sipU-lacZ*, and *sipV-lacZ* gene fusions were constructed, and introduced in the chromosome of *B. subtilis* 8G5 (schematically shown in Fig. IX.6*A*), using a similar strategy as previously described for a transcriptional *sipS-lacZ* gene fusion (33). This resulted in *B. subtilis* 8G5::pLGT207 (*sipT-lacZ*), *B. subtilis* 8G5::pLGU202 (*sipU-lacZ*), and *B. subtilis* 8G5::pLGV201 (*sipV-lacZ*), respectively. Next, these strains and *B. subtilis* 8G5::pGDE22 (*sipS-lacZ*; Ref. 33) were grown in TY and minimal medium, and samples withdrawn at hourly intervals were assayed for ß-galactosidase activity.

In both media, nearly identical results were obtained with *B. subtilis* 8G5::pLGT207 (*sipT-lacZ*): the levels of ß-galactosidase activity increased after the cells entered the transition state (t=0) between the exponential and the post-exponential growth phase (Fig. IX.6, *B* and *C*; indicated with the symbol "■"), and they continued to increase during the post-exponential growth phase, indicating that the promoter(s) of *sipT* became more active than in the exponential growth phase. Thus, the transcription of the *sipT* (Bsu) gene appears to be temporally controlled, similar to that of the *sipS* (Bsu) gene (Fig. IX.6, *B* and *C*; indicated with

the symbol "□"). In particular, in minimal medium the ß-galactosidase levels observed in the strains with the *sipS-lacZ* or *sipT-lacZ* gene fusions were comparable (Fig. IX.6*C*). In TY medium, however, the *sipS* promoter activity appeared to be 1.5- to 2-fold higher than that of *sipT* (Fig. IX.6*B*).

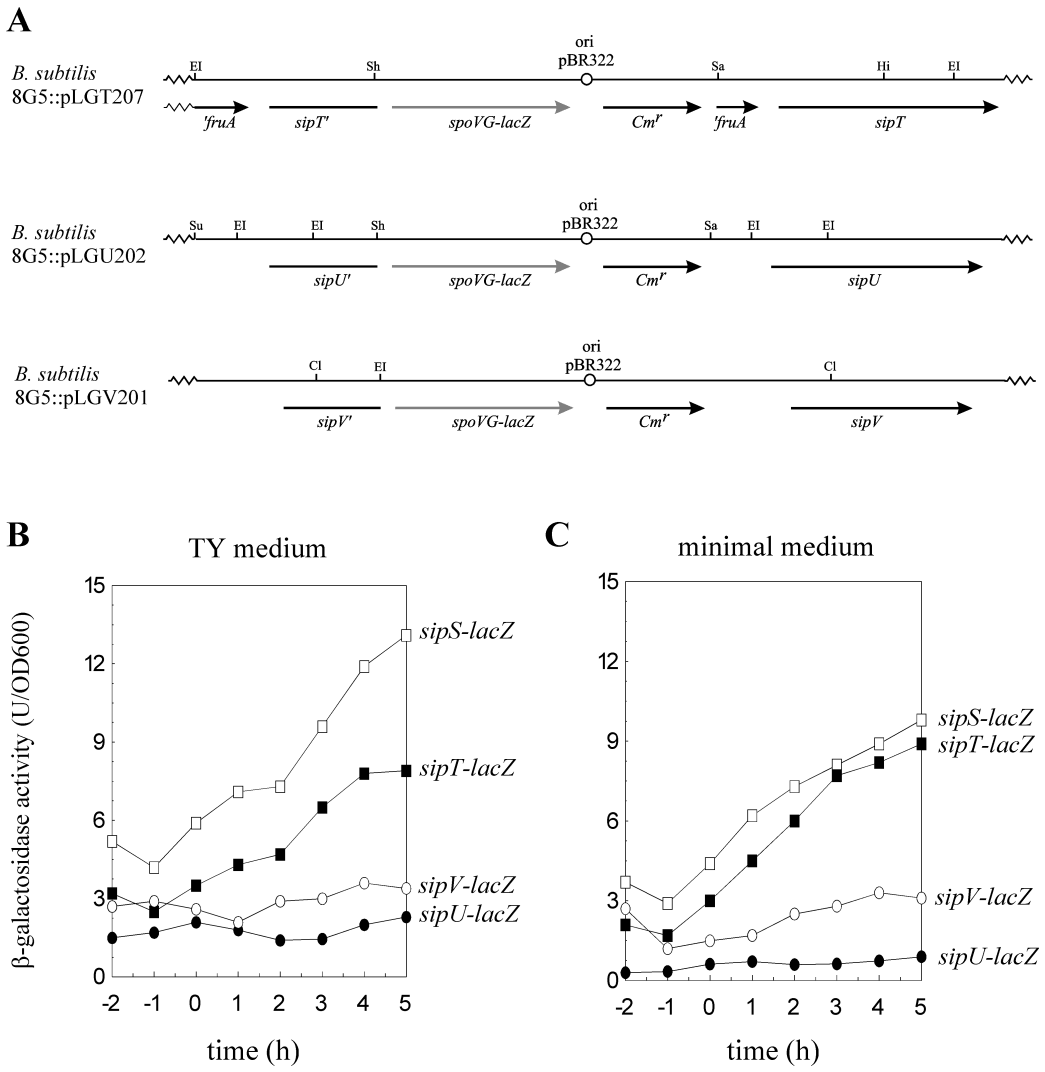


Fig. IX.6. Analysis of the expression of *sipT*, *sipU*, and *sipV* with transcriptional *lacZ* gene fusions. ***A***, Schematic presentation of the *sipT*, *sipU*, and *sipV* regions on the chromosomes of *B. subtilis* 8G5::pLGT207 (*sipT-lacZ*), *B. subtilis* 8G5::pLGU202 (*sipU-lacZ*), and *B. subtilis* 8G5::pLGV201 (*sipV-lacZ*), respectively. All three *lacZ* fusions were constructed with plasmid pLGW200 (38), a chromosomal integration plasmid for *B. subtilis* containing a promoterless *spoVG-lacZ* gene fusion. To construct a *sipT-lacZ* gene fusion, the 5' end of *sipT*, amplified by PCR with the primers lbt-9 (AT**GAATTCAGCCCGGTTATCTCC**) and lbt-12 (Fig. 1), was cloned in the multiple cloning site (MCS) upstream of the *spoVG-lacZ* fusion of pLGW200, resulting in pLGT207. Similarly, a *sipU-lacZ* gene fusion was constructed by cloning the 5' end of *sipU*, PCR-amplified with the primers lbu-3 (AGCTGTC**GACATTGCCGGACAGGCC**) and lbu-4 (AATAGG**TACCGGAGGGAACCT-CAACTTCG**), in the MCS of pLGW200, resulting in pLGU202. To construct a *sipV-lacZ* gene fusion, the 5' end of *sipV* was PCR-amplified with the primers uni (**GTAAAACGACGGCCAGT**) and lbv-2 (TTGGAATTC**GATTATCTCCAACGAC**) from a pUC18-derived plasmid carrying the *sipV* gene. The amplified fragment was cloned in the MCS of pLGW200, resulting in pLGV201. The *sip-lacZ* gene fusions were introduced in the chromosome of *B. subtilis* 8G5 by Campbell-type integration. (*Continued on next page.*)

(*Legend to Fig. IX.6 continued*)
In *B. subtilis* 8G5::pLGT207, the transcription of *lacZ* is directed by the promoter(s) of *sipT*, in *B. subtilis* 8G5::pLGU202 by the promoter(s) of *sipU*, and in *B. subtilis* 8G5::pLGV201 by the promoter(s) of *sipV*. Only restriction sites relevant for the constructions are shown (EI, *Eco*RI; Hi, *Hin*dII; Sa, *Sal*I; Sh, *Sph*I; Su, *Stu*I; Cl, *Cla*I), ori pBR322, replication functions of pBR322. ***B*** and ***C***, Time courses of the expression of *sip-lacZ* gene fusions were determined in cells growing in TY medium (***B***) or minimal medium (***C***) at 37°C. Beta-galactosidase activities (in Units per OD600) were determined for *B. subtilis* 8G5::pGDE22 (□, *sipS-lacZ*), *B. subtilis* 8G5::pLGT207 (■, *sipT-lacZ*), *B. subtilis* 8G5::pLGU202 (●, *sipU-lacZ*), and *B. subtilis* 8G5::pLGV201 (○, *sipV-lacZ*). Zero time (t=0) indicates the transition point between the exponential and the post-exponential growth phases.

The regulation of the transcription of the *sipU* and *sipV* genes seems to be completely different from that of *sipS* and *sipT*. When grown in TY or minimal medium, a nearly constant low level of ß-galactosidase activity was observed in cells of *B. subtilis* 8G5::pLGU202 (*sipU-lacZ*) and *B. subtilis* 8G5::pLGV201 (*sipV-lacZ),* irrespective of the growth phase (Fig. IX.6, *B* and *C*; indicated with the symbol "●" and"o", respectively). In fact, in both media the ß-galactosidase levels of *B. subtilis* 8G5::pLGU202 (*sipU-lacZ)* were nearly equal to background ß-galactosidase levels of control cells lacking a copy of *lacZ* (data not shown). Nevertheless, expression of the *sipU* gene was evident as colonies of *B. subtilis* 8G5::pLGU202 (*sipU-lacZ*) were blue on TY, or minimal plates with X-gal; control cells lacking the *lacZ* gene, or containing a fusion between a non-transcribed gene and *lacZ* remained white (data not shown). These findings indicate that the activity of the *sipU* and *sipV* promoter(s) does not depend on the growth phase.

IX.5. Discussion

In those microorganims of which the genomes have been sequenced completely, at most two or three homologous type I SPases seem to be present. For example, the cyanobacterium *Synechocystys* contains two type I SPases (GenBank accessions D90899 and D90904), and the yeast *Saccharomyces cerevisiae* contains three of these enzymes (52). In the latter case, these enzymes are localized in two distinct membrane systems, the inner mitochondrial membrane (*ie.* Imp1p and Imp2p; Refs. 22, and 28), and the ER membrane (*ie.* the Sec11 protein; Ref. 32). Two homologues of bacterial type I SPases are also commonly found in the ER SPase complex of higher eukaryotes (8; M.O. Lively and S.J. Walker, personal communication). In contrast, for other genetically well-characterized microorganisms, such as *E. coli* (GenBank accession ECOLI U00096), *H. influenzae* (26) and *Methanococcus jannaschii* (53), only one type I SPase seems to be sufficient, and type I SPases may even be completely absent from *Mycoplasma genitalium* (54). In our present studies we show that *B. subtilis* contains at least four chromosomally-encoded type I SPases (SipS, SipT, SipU, and SipV) involved in protein secretion. In addition, we have previously shown that certain strains of *B. subtilis* also contain plasmids (pTA1015/pTA1040) specifying related type I SPases (17). Thus, the composition of the protein secretion machinery of *B. subtilis* seems to be unique with respect to the high number of SPases.

A second remarkable property of the secretion machinery of *B. subtilis* concerns the high degree of similarity between the substrate specificities of SipS, SipT, SipU, SipV, SipP (pTA1015) and SipP (pTA1040). The conclusion that the substrate specificities of these six type I SPases are very similar is based on our present and previous (6, 17) observations that all these enzymes are able to cleave the same substrate, pre(A13i)-ß-lactamase, albeit with different efficiencies, and under different conditions. By contrast, the type I SPases Imp1p and Imp2p in the inner mitochondrial membrane seem to have completely distinct substrate specificities (28).

Though similar, the substrate specificities of the four chromosomally-encoded type I SPases of *B. subtilis* are not identical, as our present results indicate that these enzymes have, at least *in vivo*, a different preference for the precursor of the α-amylase AmyQ of *B. amyloliquefaciens*. Pre-AmyQ processing was significantly reduced in strains lacking SipT, indicating that this precursor is a preferred substrate of SipT. In contrast, SipV did not seem to be involved in pre-AmyQ processing, whereas the presence of SipS and SipU interfered with efficient processing of this precursor. Taken together, these findings suggest that SipS and SipU somehow compete with SipT for binding of pre-AmyQ, and that SipT, but not SipS and SipU, can cleave this precursor efficiently. Similarly, preliminary data suggest that pro-OmpA of *E. coli* may be a preferred substrate of SipT (A. Matzen and R. Freudl, personal communication), whereas the presence of SipT seems to interfere with efficient secretion of levansucrase of *B. subtilis* (our unpublished results).

What could be the advantage(s) for an organism, like *B. subtilis*, to acquire and maintain so many different SPase-encoding genes during its evolution? Our present observations indicate that multiple SPases may serve to guarantee a sufficient capacity for protein secretion under varioius conditions. First, we show that none of the four chromosomally-encoded type I SPases described in this manuscript is, by itself, essential for cell growth and protein secretion. As SPase activity is essential for the viability of *B. subtilis* (our unpublished results), our present observations imply that the secretory precursor processing machinery of this organism is functionally redundant. Thus, *B. subtilis* can allways avail of a "backup SPase", even in the case that a complete SPase-encoding gene would be lost. This may be of particular importance for *B. subtilis* and related bacilli, such as *B. amyloliquefaciens*, which secrete large amounts of proteins into the medium. Second, our present observations suggest that different chromosomally-encoded type I SPases of *B. subtilis* serve different functions in the exponential and post-exponential growth phases. For example, it seems likely that SipU and SipV are involved in the processing of secretory pre-proteins that are synthesized during all growth phases, because the corresponding genes are transcribed at a constitutive (low) level. In contrast, the transcription of the *sipS* and *sipT* genes is temporally regulated, the highest levels of transcription being observed in the post-exponential growth phase. The increase in the levels of transcription of *sipS* and *sipT* starts in the transition phase between the exponential and the post-exponential growth phase and is, thus, concerted with the onset of the transcription of most secretory proteins of *B. subtilis* (55). In fact, in minimal medium, the transcription of

both the *sipS* gene (33) and the *sipT* gene, but not the *sipU* and *sipV* genes, is controlled by the DegS-DegU two-component regulatory system (our unpublished results), which is also required for the transcription of several genes for secretory proteins (56). Therefore, it seems likely that SipS and SipT serve to increase the capacity for protein secretion in the post-exponential growth phase under conditions of increased synthesis of secretory proteins in *B. subtilis*. The latter hypothesis would be consistent with two of our previous findings: a), the availability of SPase can be a limiting factor for the secretion of certain hybrid precursor proteins, which can be overcome by SPase overproduction (6, 42); and b), certain endogenous plasmids of *B. subtilis* contain SPase-encoding genes, suggesting that SPase can also be a limiting factor for protein secretion in a natural system (17). In addition, the special importance of SipS and SipT for protein secretion in *B. subtilis* is underscored by our recent (unpublished) observation that only cells lacking both SipS and SipT were not viable, whereas all other *sip* gene mutations could be combined.

Finally, how many type I SPases does *B. subtilis* contain exactly? The systematic sequence analysis of the *B. subtilis* genome has been completed very recently, and it seems that there are no other genes for close homologues of SipS, SipT, SipU, and SipV (I. Moszer, personal communication). However, our computer-assisted analyses revealed one additional gene (*yqhE*) for a potential type I SPase (SipW) that is more closely related to the type I SPases from *archaea* and the eukaryotic ER membrane than to bacterial type I SPases. The question whether SipW is actively involved in protein secretion remains to be answered.

Abbreviations

The abbrevations used are: Bam, *Bacillus amyloliquefaciens*; Bsu, *Bacillus subtilis*; ER, endoplasmic reticulum; Imp, inner membrane protease; Lep, leader peptidase; SPase, signal peptidase.

References

1. Von Heijne, G. (1990) *J. Membrane Biol.* **115**, 195-201
2. Pugsley, A.P. (1993) *Microbiol. Rev.* **57,** 50-108
3. Von Heijne, G. (1994) in *Signal peptidases* (von Heijne, G., ed). pp. 1-3, R.G. Landes Company, Austin, T. X.
4. Schatz, G., and Dobberstein, B. (1996) *Science* **271,** 1519-1526
5. Dalbey, R. E., and von Heijne, G. (1992) *Trends. Biochem. Sci.* **17**, 474-478
6. Van Dijl, J. M., de Jong, A., Vehmaanperä, J., Venema, G. and Bron, S. (1992) *EMBO J.* **11,** 2819-2828
7. Dalbey, R. E. (1994) in *Signal peptidases* (von Heijne, G., ed). pp. 5-15, R.G. Landes Company, Austin, T. X.
8. Lively, M.O., and Shelness, G.S. (1994) in *Signal peptidases* (von Heijne, G., ed). pp. 59-71, R.G. Landes Company, Austin, T. X.
9. Black, M. T. (1993) *J. Bacteriol.* **175**, 4957-4961
10. Tschantz, W. R., Sung, M., Delgado-Partin, V. M., and Dalbey, R. E. (1993) *J. Biol. Chem.* **268,** 27349-27354

11. Van Dijl, J. M., de Jong, A., Venema, G. and Bron, S. (1995) *J. Biol. Chem.* **270,** 3611-3618
12. Paetzel, M. and Dalbey, R. E. (1997) *Trends. Biochem. Sci.* **22,** 28-31
13. Kuo, D. W., Chan, H. K., Wilson, C. J., Griffin, P. R., Williams, H., and Knight, W.B. (1993) *Arch. Biochem. Biophys.* **303,** 274-280
14. Pratje, E., Esser, K. H., and Michaelis, G. (1994) in *Signal peptidases* (von Heijne, G., ed). pp. 105-112, R.G. Landes Company, Austin, T. X.
15. Shelness, G. S., Lin, L., and Nicchitta, C. V. (1993) *J. Biol. Chem*. **268,** 5201-5208
16. Hoang, V., and Hofemeister J. (1995) *Biochim. Biophys. Acta* **1269,** 64-68
17. Meijer, W. J. J., de Jong, A., Wisman, G. B. A., Tjalsma, H., Venema, G., Bron, S., and van Dijl, J.M. (1995) *Mol. Microbiol.* **17,** 621-631
18. Cregg, K. M., Wilding, E. I., and Black, M.T. (1996) *J. Bacteriol.* **178,** 5712-5718
19. Philipp, W. J., Poulet, S., Eiglmeier, K., Pascopella, L., Balasubramanian, V., Heym, B.,Bergh, S., Bloom, B. R., Jacobs, W. R., and Cole, S. T. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93,** 3132-3137
20. Packer, J. C. L., André, D., and Howe, C. J. (1995) *Plant Mol. Biol.* **27,** 199-204
21. Müller, P., Ahrens, K., Keller, T., and Klaucke, A. (1995) *Mol. Microbiol.* **18,** 831-840
22. Behrens, M., Michaelis, G., and Pratje, E. (1991) *Mol. Gen. Genet.* **228**, 167-176
23. Wolfe, P. B., Wickner, W., and Goodman, J.M. (1983) *J. Biol. Chem.* **258**, 12073-12080
24. Black, M. T., Munn, J. G. R., and Allsop, A. (1992) *Biochem. J.* **282,** 539-543
25. Van Dijl, J. M., van den Bergh, R., Reversma, T., Smith, H., Bron, S., and Venema, G. (1990) *Mol. Gen. Genet.* **223**, 233-240
26. Fleischmann, R. D., Adams, M. D., White, O., and 37 other authors (1995) *Science* **269,** 496-512
27. Klug, G., Jäger, A., Heck, C., and Rauhut, R. (1997) *Mol. Gen. Genet.* **253**, 666-673
28. Nunnari, J., Fox, T. D., and Walter, P. (1993) *Science* **262,** 1997-2004
29. Date, T. (1983) *J. Bacteriol.* **154,** 76-83
30. Dalbey, R. E., and Wickner, W. (1985) *J. Biol. Chem.* **260**, 15925-15931
31. Van Dijl, J. M., de Jong, A., Smith, H., Bron, S. and Venema, G. (1988) *Mol. Gen. Genet.* **214,** 55-61
32. Böhni, P. C., Deshaies, R. J., and Schekman, R. W. (1988) *J. Cell Biol.* **106**, 1035-1042
33. Bolhuis, A., Sorokin, A., Azevedo, V., Ehrlich, S. D., Braun, P. G., de Jong, A., Venema, G., Bron, S., and van Dijl, J. M. (1996) *Mol. Microbiol.* **22,** 605-618
34. Akagawa, E., Kurita, K., Sugawara, T., Nakamura, K., Kasahara, Y., Ogasawara, N., and Yamane, K. (1995) *Microbiology* **141,** 3241-3245
35. Sorokin, A., Zumstein, E., Azevedo, V., Ehrlich, S. D., and Seror, P. (1993) *Mol. Microbiol.* **10,** 385-395
36. Leenhouts, K., Buist, G., Bolhuis, A., ten Berge, A., Kiel, J., Mierau, I., Dabrowska, M., Venema, G., and Kok, J. (1996) *Mol. Gen. Genet.* **253,** 217-224
37. Palva, I. (1982) *Gene* **19,** 81-87
38. Van Sinderen, D., Withoff, S., Boels, H., and Venema, G. (1990) *Mol. Gen. Genet*. **224,** 396-404
39. Wertman, K. F., Wyman, A. R., and Botstein, D. (1986) *Gene* **49**, 253-262
40. Bron, S. and Venema, G. (1972) *Mutat. Res.* **15,** 1-10
41. Spizizen, J. (1985) *Proc. Natl. Acad. Sci. USA* **44,** 1072-1078
42. Van Dijl, J. M., de Jong, A., Smith, H., Bron, S. and Venema, G. (1991a) *Mol. Gen. Genet.* **227,** 40-48
43. Van Dijl, J. M., de Jong, A., Smith, H., Bron, S. and Venema, G. (1991b) *J. Gen. Microbiol.* **137,** 2073-2083
44. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.
45. Sanger, F., Nicklen, S, & Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463-5467.
46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403-410
47. Miller, J. H. (1982) *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.

48. Towbin, H., Steahelin, T., and Gordon, J. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 4350-4354
49. Gay, P., Chalumeau, H. and Steinmetz, M. (1983) *J. Bacteriol.* **153,** 1133-1137
50. Evan, G.I., Lewis, G.K., Ramsay, G., and Bishop, M. (1985) *Mol Cell Biol*. **5**, 3610-3616.
51. Kontinen, V. P., and Sarvas, M. (1988) *J. Gen. Microbiol.* **134,** 2333-2344
52. Oliver, S. G. (1996) *Nature* **379,** 597-600
53. Bult, C. J., White, O., Olsen, G. J., and 37 other authors (1996) *Science* **273**, 1058-1073
54. Fraser, C. M., Gocayne, J. D., White, O., and 26 other authors (1995) *Science* **270**, 397-403
55. Ferrari, E., Jarnagin, A.S, and Schmidt, B.F. (1993) in *Bacillus subtilis and other Gram-positive bacteria* (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 917-937, American Society for Microbiology, Washington, D. C.
56. Msadek, T., Kunst, F., and Rapoport, G. (1993) in *Bacillus subtilis and other Gram-positive bacteria* (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 729-745, American Society for Microbiology, Washington, D. C.
57. Von Heijne, G. (1992) *J. Mol. Biol.* **225,** 487-494
58. Smith, H., Bron, S., van Ee, J., and Venema, G. (1987) *J. Bacteriol.* **169,** 3321-3328

CHAPTER X

Summary and conclusions

Genomics concerns the aquisition of knowledge of structure and function of genomes. The scope of genomics research is wide; it includes the determination of the nucleotide sequence of all of an organisms' DNA, the analysis of the information that resides in it, the assessment of the functions of the uncovered information and how these functions interact, and the study of how and why genomes have evolved the way they did.

The first prerequisite for genomics research is the availability of genome sequences, preferably as complete as possible. The first genome sequence, that of the Gram-negative pathogenic bacterium *Haemophilus influenzae*, became available in 1995. It has been determined using the method called whole-genome random sequencing. The most important aspect of this strategy is that the genome sequence is determined and assembled from randomly taken fragments; genetic information of the organism is not needed in advance. Now, within an interval of just three years, already 19 complete genome sequences are available, including two eukaryal genomes; those of *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. Within the next decade, many of the most important model organisms used in biological research will have been determined, including the human genome which consists of no less then 3×10^9 bp in the haploid state.

The *Bacillus subtilis* genome sequencing project was undertaken for several reasons. This bacterium has already been an important subject of scientific studies for several decades, serving as a model organism in genetics research, in particular for Gram-positive bacteria, because it exhibits two "primitive" developmental processes. These are its ability to form specialised cells (endospores), and its natural capacity to import DNA from the surrounding medium (competence) and, if homologous DNA is taken up, to incorporate this into its own chromosome. Furthermore, *Bacillus* species are widely used in industry for the production of secreted enzymes such as proteases and amylases.

The *B. subtilis* genome sequence was determined in a truly international effort, by a consortium of over 25 research groups from Europe and Japan. A particular region from the genetic map of the *B. subtilis* chromosome was assigned to each research group involved in the project. Our group was initially responsible for sequencing the chromosomal region between the genetic markers *glyB* and *addAB*, estimated to contain 96 kb of DNA. Relatively large chromosomal clones from this region were obtained through various strategies: screening of a genome bank in bacteriophage lambda, plasmid walking, long-range and

inversed long-range polymerase chain reaction (LR PCR & i-LR PCR). Of these, the two PCR applications proved to be the most efficient, since they circumvent the -often troublesome- cloning of large fragments of foreign DNA in *Escherichia coli*. The sequence of chromosomal clones was subsequently determined with the aid of one of several random shotgun approaches, of which the DNaseI shotgun method was proven to be the most successful.

In total, 171,812 basepairs (bps) were determined in our group, between 83° and 97° on the circular map of the chromosome, representing 4.1% of the *B. subtilis* genome. Due to the erratic original genetic map of this chromosomal region, this was almost twice as much as was originally foreseen at the onset of the project. We have identified 170 putative genes in this region, 48 of which could not be assigned a possible function on the basis of their amino acid similarity to known proteins in the public databases. Analyses of the deduced protein sequences with respect to compartmentalisation signals revealed that 45% of these were putative membrane proteins, 4% secreted proteins, 49% cytoplasmatic, and 2% membrane-associated proteins through lipomodification. Many of the putative genes in this region have one or more paralogs in the *B. subtilis* genome, and some of these are members of large paralog families, such as the family of ABC-type transporters and that of DegU-type transcriptional regulators. The sequencing and subsequent *in silico* analysis of the region *prkA* to *addAB* are described in Chapters II and III. The complete genome sequence of *B. subtilis* is presented in Chapter IV.

Chapter five is an example of *in silico* genomics at the proteome level. It deals with the positional analysis of amino acid (aa) frequencies in the deduced proteins of fourteen complete genomes (proteomes). This has revealed a biased use of many aa's in the amino-terminal and carboxy-terminal aa positions of these proteomes. Further investigation of these biases with respect to charge- and hydrophobicity characteristics, showed that all proteomes display similar differences between the amino- and carboxy-terminus with respect to these parameters. This could reflect that the N- and C-termini of proteins are usually located at the surface of proteins, and might be involved in the proper translocation of the nascent proteins through the ribosome.

Several chapters in this thesis deal with typical examples of functional genomics.

Chapter VI is an example of a straightforward approach. The deduced protein sequence of *yhxB*, a gene from within the *prkA* to *addAB* region, was found to be highly similar to phosphoglucomutases from other organisms. From previous studies it was known that a genetic marker encoding such a function, and being involved in the glucosylation of the cell wall component teichoic acid, resides in this chromosomal region. We have demonstrated that *yhxB* encodes the protein responsible for glucosylation of teichoic acid, that this gene has no functional paralog in the *B. subtilis* genome, and that inactivation of the gene rendered the cells resistant to several bacteriophages.

Chapter VII represents another approach to functional genomics. This chapter deals with the search for a biological function of an ubiquitous gene, *hit*, that has orthologs in

probably all living organisms. A strain in which this gene was inactivated was subjected to systematic functional analyses, which included analysis of transcription, regulation, biochemistry, and phenotype screening. The Hit protein was found to convert ADP in two ways, the relative amounts of products of the Hit-mediated reaction being dependent on the pH. Hit hydrolysed ADP to AMP and Pi, and also acted as phosphotransferase in the reaction 2 ADP $\rightarrow$ ATP + AMP. Phenotype screening revealed that the gene was involved in heat-shock protection in *B. subtilis*. Another ubiquitous gene, *yabJ*, was observed to be involved in the regulation of transcription of the *hit* gene.

In Chapter VIII, the identification of a new forespore-specific gene of *B. subtilis* is described. Based on the observed presence of two aa sequence motifs in the deduced protein, specific for small acid-soluble spore proteins (SASP's) and membrane-anchored lipoproteins, we assumed that this gene might be associated with the sporulation process. This assumption has subsequently been validated. YhcN was shown to be localised in the inner spore membrane, and inactivation of the gene yielded a strain that was impaired in spore germination.

Chapter IX deals with a typical example of paralog research. The *B. subtilis* chromosome encodes four paralogous type I signal peptidases, responsible for the removal of signal peptides from secretory precursor proteins. Functional analysis of these signal peptidase genes and the corresponding proteins revealed that the latter have similar but non-identical substrate specificities, and that the genes have different expression characteristics. The *sipU* and *sipV* genes are transcribed constitutively at a low level, while transcription of *sipS* and *sipT* is temporally controlled.

The availability of complete genome sequences has already drastically changed the way in which genetic research is performed. Until recently, when a particular function of an organism was investigated, a researcher first had to clone the corresponding gene with the aid of various time-consuming strategies. Today, in the genomics era, a variety of techniques exists to identify the particular gene in a genome sequence that encodes the function one is looking for. However, the real added value of genomics lies in several other aspects. First, it is now feasible to systematically analyse all the (unknown) genes from a genome in either a simultaneous, or a high-throughput serial approach. Secondly, it now becomes possible to investigate the regulation of all genes in a genome at once, as well as the interactions between the encoded proteins and the interactions between the proteins and the genes. The complete genome sequences also enable researchers to investigate evolutionary relationships between organisms in a new and exciting way. Finally, major advantages from genomic research are to be expected in the field of biotechnological and medical application.

Samenvatting en conclusies

Genomica omvat het onderzoek aan structuur en functie van genomen. Het werkveld van de genomica is breed; het omvat de bepaling van de basenvolgorde van het DNA van een organisme, de analyse van informatie die daarin besloten ligt, de functiebepaling daarvan, de analyse van de interactie tussen de gevonden informatie-eenheden en de studie van genoom-evolutie.

De eerste voorwaarde voor het bedrijven van genoom-onderzoek is de beschikbaarheid van, bij voorkeur complete, genoomvolgordes. De eerste genoomvolgorde, die van de Gram-negatieve pathogene bacterie *Haemophilus influenzae*, kwam beschikbaar in 1995. Deze werd bepaald door middel van de methode “whole-genome random sequencing”. Het belangrijkste aspect van deze strategie is dat de basenvolgorde van willekeurige DNA fragmenten bepaald wordt; genetische informatie over het organisme is niet nodig. Nu, slechts iets meer dan drie jaar later, zijn al 19 complete genoomvolgordes bekend, waaronder de eukaryote genomen van *Saccharomyces cerevisiae* en *Caenorhabditis elegans*. Binnen het komende decennium zullen vele van de belangrijkste model organismen voor biologisch onderzoek bekend zijn op nucleotide niveau, waaronder het in haploide vorm niet minder dan 3×10^9 basenparen tellende genoom van de mens.

Het project voor de bepaling van de genoomvolgorde van *Bacillus subtilis* werd ondernomen om meerdere redenen. Deze bacterie, die model staat voor Gram-positieve bacteriën, is al meerdere decennia een belangrijk onderwerp van wetenschappelijk onderzoek, niet in het minst omdat het twee “primitieve” ontwikkelingssystemen kent, namelijk het vermogen om gespecialiseerde cellen (endosporen) te vormen, en het natuurlijke vermogen om DNA uit het omringende medium op te nemen (competentie) en, als het soortseigen DNA betreft, dit vervolgens in het eigen chromosoom te incorporeren. Verder worden verschillende *Bacillus* soorten veelvuldig in de industrie gebruikt voor de produktie van gesecreteerde enzymen, zoals proteasen en amylasen. De sequentie van het *B. subtilis* genoom werd in een geïntegreerd international samenwerkingsverband bepaald door een consortium van meer dan 25 onderzoeksgroepen uit Europa en Japan. Iedere participerende onderzoeksgroep kreeg een bepaald gebied op de genetische kaart van het *B. subtilis* chromosoom toegewezen. Onze groep werd bij het begin van het project het gebied tussen de genetische merkers *glyB* en *addAB* toegewezen, hetgeen naar schatting 96 kb DNA zou omvatten. Relatief grote chromosomale fragmenten werden verkregen door gebruik te maken van één van de volgende strategieën: Screenen van een genomische bank in bacteriofaag lambda, “plasmid walking”, lange-afstands-PCR en omgekeerde lange-afstands-PCR (LR PCR & i-LR PCR). De twee PCR toepassingen bleken het meest succesvol, omdat hiermee het -meestal problematische- kloneren van grote soortsvreemde DNA fragmenten in *E. coli* kon worden omzeild. De volgorde van chromosomale klonen werd vervolgens bepaald met behulp van één van een aantal mogelijke “shotgun” technieken, waarvan de DNaseI “shotgun” techniek het meest succesvol is gebleken.

In totaal werden door onze groep 171.812 basenparen (bps) bepaald van het gebied tussen 83° en 97° op de circulaire kaart van het chromosoom; dit komt overeen met 4,1% van het genoom van *B. subtilis*. Dit is bijna twee keer zoveel als bij het begin van het project was voorzien omdat de oorspronkelijke genetische kaart van dit chromosomale gebied grote fouten bleek te bevatten. In dit gebied werden 170 mogelijke genen geïdentificeerd, waarvan aan 48 geen mogelijke functie toegewezen kon worden op basis van de gelijkenis van de van hun sequentie afgeleide aminozuur volgorde met bekende eiwitten in de publieke databanken. Analyse van de compartimentalisatie signalen in de afgeleide aminozuur volgordes liet zien dat naar schatting 45% van de eiwitten membraaneiwit is, 4% wordt gesecreteerd, 49% cytoplasmatisch is en 2% membraan-geassocieerd is door middel van lipomodificatie. Veel van de genen in dit gebied hebben één of meerdere paralogen in het *B. subtilis* genoom en sommige van de afgeleide eiwitproducten zijn leden van een grote familie van paralogen, zoals de familie van ABC-type transporters of de familie van DegU-type transcriptionele regulatoren. De DNA volgorde bepaling en de daaropvolgende *in silico* analyses van het chromosomale gebied tussen *prkA* en *addAB* worden beschreven in de Hoofdstukken II en III. De complete genomische DNA volgorde van *B. subtilis* staat beschreven in Hoofdstuk IV.

Hoofdstuk V is een voorbeeld van *in silico* genomica op het proteoom niveau. Het behandelt de positionele analyse van aminozuur frequenties in de afgeleide eiwitvolgordes van veertien bekende genomen (proteomen). Deze analyse liet een onder-, c.q. over-representatie van vele aminozuren zien in de amino- en carboxy-terminale posities van deze proteomen. Verdere analyse van deze “biases” met betrekking tot de ladings- en hydrofobiciteits-eigenschappen liet zien dat alle proteomen vergelijkbare verschillen vertonen in deze eigenschappen tussen de amino- en carboxy-terminus. Dit kan een gevolg zijn van het feit dat N- en C- termini van eiwitten gewoonlijk aan de buitenkant gelokaliseerd zijn, maar ook van de wijze waarop het eiwitmolecuul door het ribosoom wordt getransporteerd.

Verschillende hoofdstukken in dit proefschrift behandelen typische voorbeelden van functionele genomica.

Hoofdstuk VI is een voorbeeld van een directe benadering. De afgeleide eiwitvolgorde van *yhxB*, een gen uit het *prkA-addAB* gebied, bleek grote overeenkomst te vertonen met fosfoglucomutasen uit andere organismen. Uit eerder onderzoek was bekend dat genetische informatie voor een dergelijke functie, betrokken bij de glucosylering van de celwand-component teichoïnezuur, gelegen moest zijn in dit deel van het chromosoom. Wij hebben aangetoond dat *yhxB* het eiwit specificeert dat verantwoordelijk is voor de glucosylering van teichoïnezuur, dat het geen functionele paraloog in het *B. subtilis* genoom heeft en dat de inactivatie van dit gen de cellen resistent maakt tegen een aantal bacteriofagen.

Het onderzoek beschreven in hoofdstuk VII heeft betrekking op een andere benadering in functionele genomica. Dit hoofdstuk behandelt de zoektocht naar een biologische functie van een universeel gen, *hit*, dat orthologen heeft in waarschijnlijk alle levende organismen. Een stam waarin dit gen was uitgeschakeld werd onderworpen aan systematische functionele analyses, die ondermeer de analyse van de transcriptie, de regulatie, en het zoeken naar een

fenotype en de biochemische functie inhield. Het Hit eiwit bleek ADP op twee manieren te kunnen omzetten, waarbij de onderlinge verhoudingen van de door Hit gevormde producten afhankelijk waren van de pH. Hit hydrolyseert ADP tot AMP en Pi, maar werkt ook als fosfotransferase in de reactie 2 ADP → ATP + AMP. Fenotype screening toonde aan dat het *hit* gen betrokken was bij bescherming tegen hitte-schok in *B. subtilis*. Verder bleek dat een ander, waarschijnlijk ook universeel gen, *yabJ*, betrokken was bij de transcriptie regulatie van het *hit* gen.

In hoofdstuk VIII wordt de identificatie van een nieuw voorspore-specifiek gen van *B. subtilis* beschreven. Gebaseerd op de waarneming dat de afgeleide eiwitvolgorde van dit gen twee sequentie motieven bevatte die specifiek zijn voor “small acid soluble proteins” (SASP’s) en membraan-verankerde lipo-eiwitten, werd aangenomen dat dit gen betrokken zou kunnen zijn bij het sporulatie proces. Deze aanname werd bevestigd. Het YhcN eiwit werd in de binnenste sporemembraan aangetoond en door inactivatie van het gen werd de spore-ontkieming van de corresponderende stam minder efficiënt.

Hoofdstuk IX betreft een typisch voorbeeld van paralogen onderzoek. Het *B. subtilis* chromosoom specificeert vier paralogen van type I signaal peptidasen, die verantwoordelijk zijn voor de verwijdering van signaal peptides van uitgescheiden eiwitten. De functionele analyse van deze genen en hun corresponderende eiwitten toonde aan dat deze eiwitten overeenkomstige, maar niet volstrekt identieke, substraat specificiteiten bezitten en dat de genen verschillend tot expressie komen: de genen *sipU* en *sipV* worden constitutief getranscribeerd op een laag niveau, terwijl *sipS* en *sipT* temporeel worden gereguleerd.

De beschikbaarheid van de nucleotidenvolgordes van genomen heeft de manier waarop genetisch onderzoek wordt gedaan nu reeds drastisch veranderd. Tot voor kort moest een onderzoeker, wanneer een bepaalde functie van een organisme werd onderzocht, het corresponderende gen met behulp van verschillende tijdrovende klonerings-strategieën in handen zien te krijgen. Tegenwoordig, in het genomica-tijdperk, bestaat er een verscheidenheid aan -minder tijdrovende- technieken om het specifieke gen te identificeren dat voor de functie die men zoekt verantwoordelijk is. De echte toegevoegde waarde van de genomica ligt echter in een aantal andere aspecten. Ten eerste is het nu mogelijk om systematisch alle (onbekende) genen van een genoom op een seriële “high-throughput” manier, te analyseren. Ten tweede wordt het nu mogelijk de regulatie van alle genen in een genoom gelijktijdig te onderzoeken, evenals de interacties tussen de gevormde eiwitten en de interacties tussen de eiwitten en de genen. Tevens maken complete genoomvolgordes het de onderzoeker mogelijk om evolutionaire verwantschappen tussen organismen te onderzoeken op een nieuwe, spannende, manier. Tenslotte ligt het in de verwachting dat genomisch onderzoek van groot belang zal zijn voor biotechnologische en medische toepassingen.

Chapter XI

Hoofdstuk voor de leek

XI.1. Inleiding

Waarom dit hoofdstuk voor de leek? Dat heeft twee redenen. Ten eerste omdat mij de afgelopen jaren vele malen is gevraagd waar ik nou precies mee bezig was en ik dat, helaas, slechts enkele keren duidelijk heb kunnen maken. Ten tweede omdat ik het gewoon leuk vind om te doen. Omdat het onderzoek waar ik mee bezig ben geweest moeilijk is te begrijpen zonder basiskennis, zal ik eerst een aantal principes van de erfelijkheid uiteenzetten. Wat is DNA, wat voor informatie ligt daar precies in opgeslagen en hoe wordt deze informatie vertaald in eigenschappen? Tenslotte volgt een overzicht van het wetenschapsgebied waar ik mij de afgelopen vijf jaar mee bezig heb gehouden, de genomica.

XI.2. DNA, codering, eiwitten

Iedereen die af en toe TV kijkt weet tegenwoordig wel dat DNA de basis van het leven vormt, de blauwdruk waarin alle eigenschapen van een organisme (dit is een biologische term om een levend wezen, van welke aard dan ook, mee aan te duiden) zijn vastgelegd. DNA, waarvan de volledige Engelse naam **D**eoxyribo**N**ucleic **A**cid is, bestaat uit een viertal bouwstenen, DNA basen of nucleotiden genoemd, die in een lange keten aan elkaar gekoppeld zijn. Het bijzondere aan DNA is ook nog dat zo'n keten heel nauw geassocieerd is met een tweede DNA keten die complementair aan de eerste is. De informatie-elementen van DNA zijn de basenparen (bp), bestaande uit twee tegenover elkaar staande complementaire basen. Tegenover een A staat altijd een T en tegenover een G staat altijd een C. Schematisch is dit in Fig. XI.1. weergegeven. Een chromosoom bestaat uit één DNA dubbelmolecule, bestaande uit twee complementaire ketens van aan elkaar gekoppelde basenparen. Deze

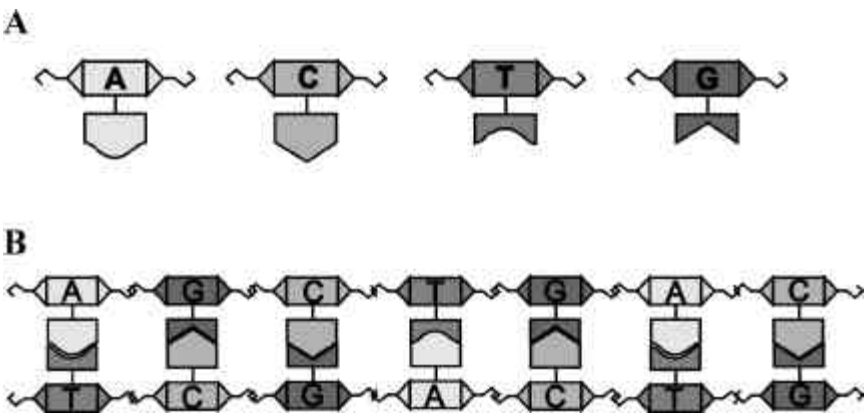


Fig. XI.1. (**A**) Schematische weergave van de vier basen van het DNA en (**B**) hoe ze in het DNA molecuul aan elkaar zijn gekoppeld en basenparen vormen. De bovenste en onderste keten vertegenwoordigen ieder een complementaire streng van de dubbele helix.

ketens kunnen uitzonderlijk lang zijn, en daarom bestaat een chromosoom bij hogere organismen niet alleen uit DNA, maar zijn er ook nog allerlei eiwitten mee geassocieerd om de juiste structuur en organisatie te handhaven. Bij de mens is al het erfelijk materiaal, hetgeen bestaat uit een totaal van zo'n drie miljard van die basenparen, verdeeld over drieëntwintig chromosomen. De bacterie waaraan ik de afgelopen tijd heb gewerkt heeft "slechts" ruim vier miljoen van die basenparen, allemaal gelegen in één circulair chromosoom. Maar hoe krijg je nu uit die code een levende en delende cel? Het is immers niet het DNA dat de biochemische functies van de cel uitvoert. We zouden hier een analogie met de computer kunnen maken. Het DNA is dan de "harde schijf" van de cel; het bevat alle informatie om cellen en organismen te kunnen laten functioneren. Als de computer een bepaalde taak gaat uitvoeren, wordt het desbetreffende programma gestart en in werking gezet. Eén programma van het chromosoom, de kleinste functionele eenheid, noemen wij een gen. Het aanschakelen van een gen houdt in dat de DNA-code van dat gen vertaald wordt in een eiwit. Eiwitten zijn ook moleculen die bestaan uit ketens van aaneengeschakelde basiseenheden, in dit geval aminozuren. Eiwitten zijn uiteindelijk de echte uitvoerders van functies in de cel en vormen tevens de belangrijkste bouwstenen van de cel. Hoe het vertalen van DNA in eiwit in zijn werk gaat is weergegeven in de Figuren XI.2 & XI.3.

```
  H  T  K  Q  F  H  K  E  E  P  S  C  I  V  Q  R  I  V   vertaling 1 heen
   I  L  N  N  F  I  R  R  N  P  H  A  L  C  R  E  L     vertaling 2 heen
 T  Y  X  T  I  S  X  G  G  T  L  M  H  C  A  E  N  C    vertaling 3 heen
ACATACTAAACAATTTCATAAGGAGGAACCCTCATGCATTGTGCAGAGAATTGTA  DNA streng 1 (F)
TGTATGATTTGTTAAAGTATTCCTCCTTGGGAGTACGTACACGTCTCTTAACAT  DNA streng 2 (R)
  M  S  F  L  K  M  L  L  F  G  X  A  N  H  L  S  N  Y   vertaling 1 terug
 C  V  L  C  N  X  L  S  S  G  E  H  M  T  C  L  I  T    vertaling 2 terug
   Y  X  V  I  E  Y  P  P  V  R  M  C  Q  A  S  F  Q     vertaling 3 terug
```

Fig. XI.2. Weergave van een DNA fragment en de zes mogelijke vertalingen daarvan in aminozuur volgordes, weergegeven in 1-letter code. Twee stopcodons (TAA & TGA; X = geen aminozuur), twee startcodons (ATG; codeert voor M = methionine) en een hypothetisch gen-product (vertaling 2 terug; MHEGSSL) zijn vet gedrukt. Zie de tekst voor verdere uitleg.

In Fig. XI.2. is weergegeven hoe de genetische code in het DNA besloten ligt. In het midden van deze figuur staat de basenvolgorde van een DNA-fragment (of DNA-sequentie) weergegeven. Te zien is dat er twee DNA volgordes zijn, die elkaars complement vormen. De code die in het DNA besloten ligt is als volgt geordend. Eiwitten, die in de cel alle biochemische functies uitvoeren en grotendeels de bouwstenen voor de cel zijn, bestaan uit ketens van enkele tientallen tot enkele duizenden aminozuren die, net als de nucleotiden van het DNA, achter elkaar zijn gekoppeld. Er bestaan twintig soorten aminozuren met verschillende biochemische karakteristieken en de aminozuur-volgorde in het eiwit bepaalt de eigenschappen daarvan. Om de twintig verschillende aminozuren in het DNA te kunnen coderen zijn er dus minimaal twintig codes, of woorden, nodig om ze te beschrijven. Een rekensommetje leert dat er, met gebruikmaking van de vier letters G, A, T, en C, woorden van minimaal drie letters nodig zijn om ze allemaal te kunnen beschrijven. Met twee letters zouden slechts $4 \times 4 = 16$ woorden gevormd kunnen worden. Echter, met 3 letter kunnen $4 \times$

$4 \times 4 = 64$ woorden van drie letters gevormd worden. En dat is precies wat de natuur gebruikt. Een woord van drie letters dat een aminozuur specificeert heet een *codon*. Omdat er een overschot aan mogelijke codons ten opzichte van het aantal aminozuren is (64 mogelijkheden voor 20 aminozuren), bestaat er een zogenaamde redundantie. Elk aminozuur kan door een aantal verschillende codons gespecificeerd worden. Zo is in Fig. XI.2. te zien dat in "vertaling 3 heen" van het DNA de "T", welke staat voor aminozuur threonine, gecodeerd wordt door het codon ACA, maar ook door het codon ACC. In de DNA-volgorde moet natuurlijk aangegeven worden waar de start van de beschrijving van een eiwit ligt en waar het einde is. De start- en stopcodons vormen deze signalen. Er verschillende varianten van elk van deze. In de figuur is van elk één voorbeeld vet gedrukt aangegeven; TAA is zo'n stopcodon (X in de vertaling; er wordt geen aminozuur ingebouwd) en ATG is een startcodon (M in de vertaling; er wordt een methionine ingebouwd). In Fig. XI.3. is weergegeven hoe de vertaling van DNA naar een aminozuur-keten (eiwit) in de cel plaatsvindt.

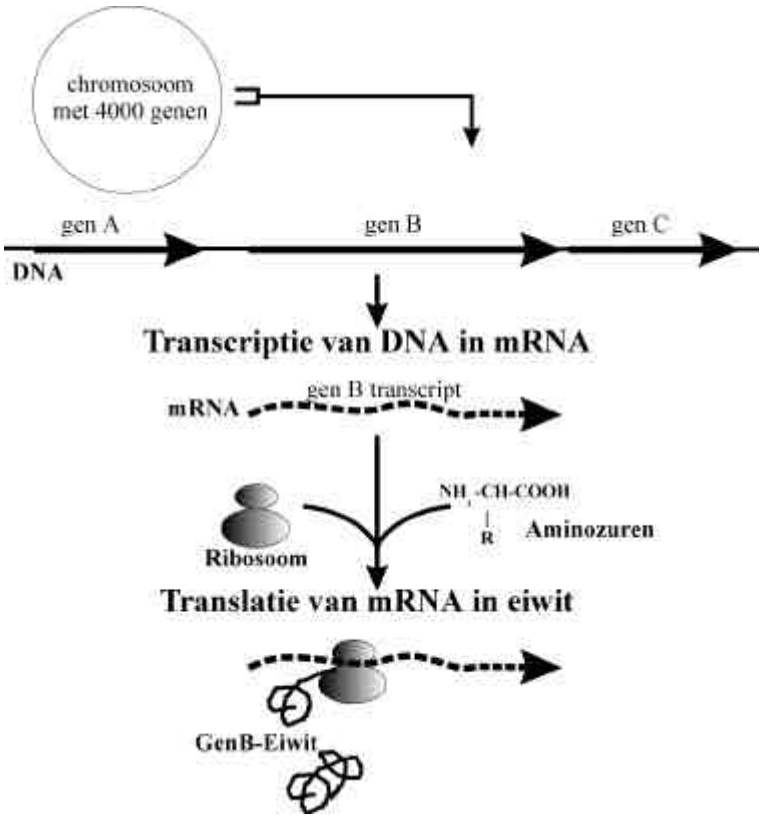


Fig. XI.3. Overzicht van het proces van het overschrijven van het DNA van een hypothetisch gen B in mRNA (transcript) gevolgd door de vertaling, op het ribosoom, van dit mRNA in de aminozuur-volgorde van het eiwit dat door gen B wordt gespecificeerd. De variabele groep van het aminozuur, die het zijn specifieke eigenschappen geeft, is aangegeven met een R. Zie ook de tekst.

Van een heel klein gedeelte uit het chromosoom (Fig. XI.3., links boven) is een uitvergroting gemaakt. Drie genen zijn door middel van pijlen weergegeven. De start van een pijl komt overeen met de plaats van een startcodon en de punt met een stopcodon. Wanneer de cel behoefte heeft aan het eiwit dat door gen B wordt gespecificeerd, wordt van dit gen een kopie gemaakt in de vorm van zogenaamd boodschapper, of messenger RNA (mRNA). Dit mRNA is chemisch bijna identiek aan het DNA, maar de kopie van één zo'n gen is natuurlijk veel kleiner dan het chromosoom. Slechts die informatie die nodig voor het door gen B omschreven eiwit wordt gekopieerd in een boodschap, in een proces dat we *transcriptie* noemen. Vervolgens wordt de mRNA-boodschap naar het gedeelte van de cel gebracht waar

de vertaling ervan in eiwit plaatsvindt, het ribosoom. Het mRNA bindt aan het ribosoom, dat het als een ponskaart afleest waarbij de streng van aminozuren, zoals gespecificeerd door de DNA-volgorde wordt gevormd. Dit proces heet *translatie*. Als we uit Figuur XI.2. de vertaling "2 terug" nemen, is daar een hypothetisch gen te zien dat uit 8 codons bestaat, specificerend voor een eiwit van 7 aminozuren (vet weergegeven): Start (ATG; Methionine), Histidine (H), Glutamine (E), Glycine (G), Serine (S), Serine (S), Leucine (L), en stop (X; geen aminozuur). Alle aminozuren hebben dezelfde chemische basisstructuur, die de ruggegraat van het eiwit vormt (in Fig. XI.3 met NH_2-CH-COOH aangegeven) en een variabel gedeelte dat elk aminozuur zijn specifieke eigenschappen geeft (in Fig. XI.3 met R, Restgroep aangegeven).

XI.3. DNA replicatie en celdeling

Nu bekend is hoe de codering in elkaar zit, kunnen een aantal andere begrippen eens wat nader onder de loep genomen worden. In de vorige paragraaf is beschreven dat DNA de informatie voor leven bevat en dat deze vertaald wordt in ketens van aminozuren, de eiwitten. Eiwitten zijn de macromoleculen die de biochemische processen uitvoeren en voor de structuur van de cel zorgen. Kortom, eiwitten zijn de uitvoerders van (bijna) alle processen in de levende cel terwijl DNA de informatiedrager voor die processen is. Een paar voorbeelden ter verduidelijking. Het voedsel dat wij eten wordt afgebroken door eiwitten (enzymen genoemd, omdat ze een chemische reactie katalyseren) en onze spieren en huid bestaan voornamelijk uit eiwitten (structurele eiwitten). Ook zijn er vele eiwitten die processen reguleren, zoals het 'aanschakelen' van genen (regulator eiwitten).

Een belangrijk kenmerk van levende cellen is de mogelijkheid om zich te delen in twee genetisch identieke nakomelingen. Dit proces van celdeling, zoals dat gebeurt in bacteriën samengevat in Figuur XI.4, volgt een vast patroon. Als een cel zich gaat delen, wordt eerst een kopie van het erfelijk materiaal gemaakt. Dit gebeurt zoals aangegeven in Fig. XI.4. Eerst worden de twee strengen van het DNA dubbelmolecule van elkaar losgemaakt; de dubbele helix wordt ontwonden. Dit proces start bij de replicatie-startplaats en dit zijn speciaal hiertoe dienende gebieden op het chromosoom. Vervolgens wordt tegenover elke DNA streng een nieuwe, complementaire, DNA streng gesynthetiseerd. Dit proces heet DNA replicatie. De DNA replicatie eindigt bij de replicatie-terminatieplaatsen op het chromosoom. Uiteindelijk, als de replicatie voltooid is en er dus twee identieke dochterchromosomen zijn gemaakt, komen deze van elkaar los. Nu het erfelijk materiaal verdubbeld is kan de celdeling doorgaan die, wanneer voltooid, resulteert in de vorming van twee dochtercellen met dezelfde erfelijke informatie als die van de cel waaruit ze zijn gevormd.

XI.4. Mutaties en evolutie

De motor van het evolutieproces omvat drie componenten: mutatie, selectie en recombinatie. De eerste essentiële component van evolutie is het ontstaan van variatie, ofwel mutaties in de erfelijke informatie. De tweede component is selectie op de ontstane variatie en de laatste betreft de vorming van nieuwe combinaties met variaties in erfelijke informatie.

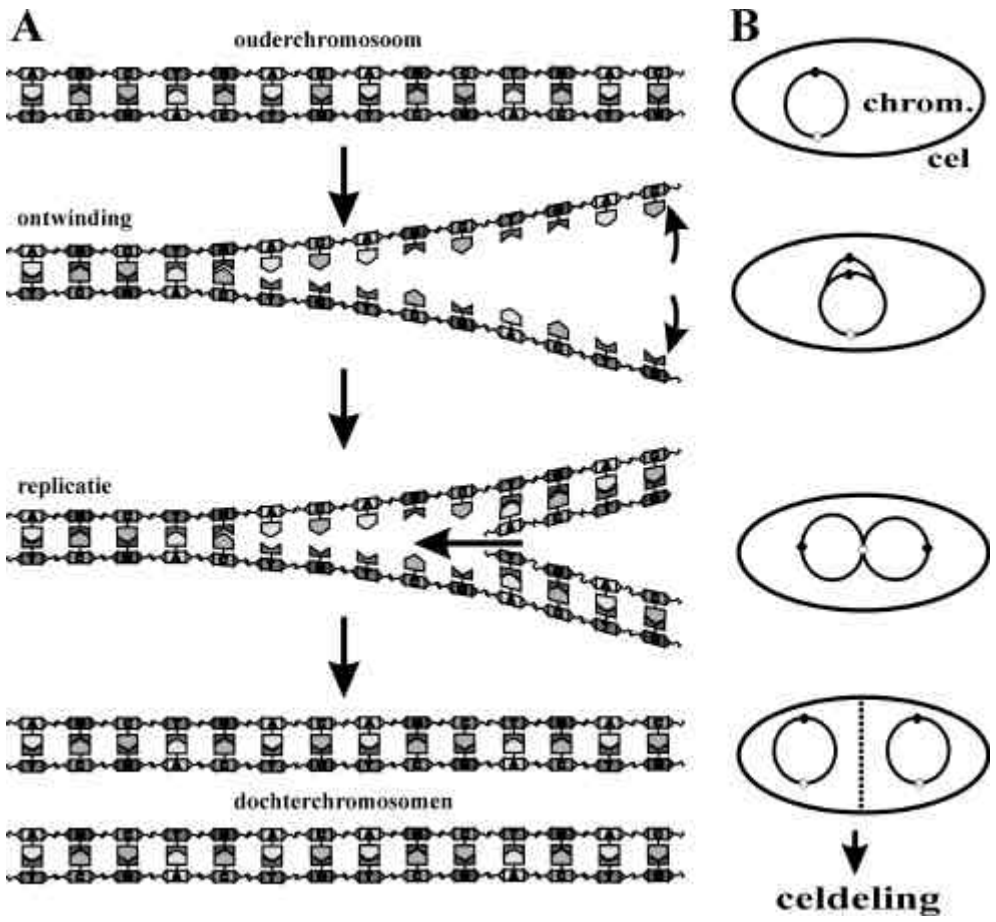


Fig.XI.4. Schematische voorstelling van DNA replicatie (A) en dit proces in de context van de bacteriële celdeling (B). In figuur B staat chrom. voor chromosoom; een gesloten cirkeltje op het chromosoom staat voor de replicatie-startplaats, en een open cirkeltje illustreert de replicatie-terminatieplaats. Zie ook de tekst.

Mutatie betekent niets meer dan verandering. Hoewel er uitgebreide controle mechanismen zijn ingebouwd, worden er tijdens het DNA replicatieproces wel eens fouten gemaakt. Ook kan het gebeuren dat DNA-schade, opgelopen door bepaalde (mutagene) stoffen of straling (bijvoorbeeld *u.v.* licht), niet op de juiste manier wordt hersteld, d.w.z. dat het herstel niet leidt tot de oorspronkelijke situatie. Tegenover een C nucleotide kan, bijvoorbeeld, in plaats van de correcte G, wel eens een T in het DNA ingebouwd worden. Daardoor kan de code voor het aminozuur op die positie veranderen. Zo zou (in Fig. XI.2.), als in de vierde base van het eerder beschreven vet gedrukte gen de base C vervangen wordt door een T, het Histidine aminozuur behorende bij die positie (gecodeerd door het CAT codon), veranderen in een Tyrosine (Y, gecodeerd door TAT). De aminozuur volgorde van het desbetreffende eiwit is dan niet meer $MHEGSSL_{stop}$, maar $MYEGSSL_{stop}$. Dit is echter zeker niet het enige type verandering dat kan plaatsvinden. Ook kunnen er nucleotiden te veel of te weinig worden ingebouwd, en die kunnen een verandering van leesframe (dit zijn de zes vertalingen uit Fig. XI.2) tot gevolg hebben. Tenslotte kunnen er hele stukken van het DNA – van een paar nucleotiden tot vele duizenden- verplaatst worden, verdwijnen of verdubbeld worden. In Fig. XI.5 staat samengevat welke mutaties er zoal kunnen plaatsvinden in het

DNA en wat de consequenties van deze mutaties zijn voor ontstaan van nieuwe (varianten van) eiwitten. Dit overzicht is echter nog niet volledig. Al deze processen zijn min of meer willekeurig en komen in elk levend wezen voor.

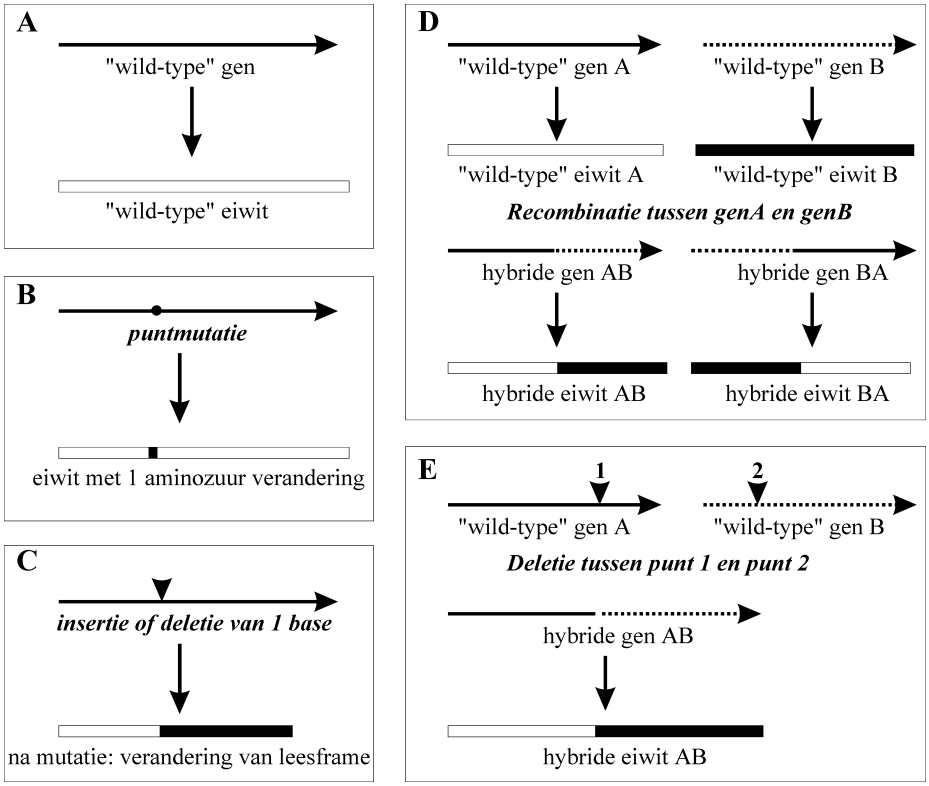


Fig.XI.5. A-E: Overzicht van enkele typen veranderingen die in het DNA kunnen optreden en de gevolgen daarvan voor het eiwit dat door het desbetreffende gen wordt gespecificeerd. Wild-type (w.t.) staat voor de oorspronkelijke situatie.

Als een mutatie is ontstaan, kan het effect hiervan schadelijk, neutraal of gunstig zijn voor de overlevingskansen en/of het reproduktiesucces van het desbetreffende organisme. Elk organisme concurreert met andere organismen om voedsel en -in het geval van de hogere organismen- een voortplantingspartner. Een organisme met een mutatie die zijn concurrentiepositie om voedsel en voortplanting minder sterk maakt, zal een kleinere kans hebben om zijn erfelijke informatie op de volgende generatie over te dragen dan een niet-gemuteerde soortgenoot.

De derde component in evolutieprocessen is recombinatie. Het bekendste mechanisme hiervoor is seksuele voortplanting. Het gevolg en doel hiervan zijn dat nieuwe combinaties van eigenschappen worden verkregen. Een voorbeeld voor die vorming van nieuwe combinaties: Een vader met bruine ogen en steil haar verwekt bij een vrouw met blauwe ogen en krullend haar een kind met bruine ogen en krullend haar. Bij hogere organismen (planten en dieren) komt seksuele voortplanting erop neer dat de helft van de genetische informatie van beide ouders in het nageslacht samengevoegd wordt. Bij lagere organismen zijn veelal andere mechanismen aanwezig voor de uitwisseling van genetische informatie, zelfs tussen verschillende soorten.

Nu de basismechanismen van het leven, voor wat betreft celdeling en voortplanting globaal bekend zijn, kunnen we proberen inzicht te krijgen in het proces van de evolutie zoals dat zich overal ter wereld afspeelt. Het beste kan dit onderwerp aan de hand van een

voorbeeld worden uitgelegd dat ik in het eerste jaar van mijn studie behandeld heb gekregen. De essentie van het verhaal is als volgt.

In Engeland kwam in de vorige eeuw een soort nachtvlinder (mot) voor die zich gedurende de dag verschool op de schors van een boom. Deze vlinder had daar een uitermate geschikte schutkleur voor. De vlinder was licht van kleur en was zeer moeilijk te onderscheiden tegen de achtergrond van de boom, die ook licht van kleur was door de daarop groeiende korstmossen. Hij was dus zeer goed aangepast om te overleven in de omgeving waarin hij voorkwam. Toen op een gegeven moment de industriële revolutie goed op gang begon te komen, volgde de luchtvervuiling met gelijke tred. Nu is het zo dat korstmossen zeer gevoelig zijn voor luchtvervuiling en de hoeveelheid korstmossen op de bomen waar de nachtvlinder zich op verschool ging in die periode drastisch achteruit. Het gevolg was dat de bomen een donkerder kleur begonnen te krijgen aangezien hun bastkleur donkerder was dan die van de korstmossen. De lichtgekleurde nachtvlinder begon op te vallen tegen de donkerdere achtergrond van de bomen en eindigde steeds regelmatiger als snack voor een vogel, of voor welk beest dan ook dat motten eet. Door een mutatie ontstond er in deze zelfde periode ook een mutant van de mot en wel eentje die een donkere vleugelkleur had (we weten nu dat dit het gevolg kan zijn van slechts één basenverandering in het gen dat verantwoordelijk is voor de pigmentaanmaak in de vleugels, dus erg onwaarschijnlijk is deze gebeurtenis niet). Omdat deze mutante mot veel minder opviel tegen de nu donkere achtergrond van de bomen, had deze een grotere overlevingskans dan zijn lichtere soortgenoten en het gevolg daarvan was dat hij een grotere kans had zich voort te planten. De rest kun je je voorstellen, ook zonder gebruikmaking van allerlei ingewikkelde formules; binnen afzienbare tijd was de populatie lichtgekleurde motten vrijwel geheel vervangen door de donkere variant.

Aan de hand van het bovenstaande voorbeeld wil ik nu proberen uit te leggen wat, behalve de basisvoorwaarden van verandering en uitwisseling van genetische informatie, belangrijke factoren in het evolutieproces op aarde zijn en hoe ze samenhangen.

Eerst was er de lichte mottensoort, goed aangepast aan zijn levensomgeving. Toen de leefomgeving veranderde, de korstmossen van de bomen verdwenen, was die aanpassing aan de leefomgeving niet meer waardevol en werden de motten opgegeten. De mot was niet veranderd en toch was de waarde van zijn overlevings-‘arsenaal’ minder waardevol geworden! Vervolgens verscheen de donkere variant, die met de verandering van kleur zijn overlevingskans aanzienlijk vergrootte. Kun je nu een waardeoordeel aan de kleur geven? Kennelijk is dit alleen mogelijk als je het probleem bekijkt in de context van alle (omgevings) factoren die van invloed zijn op de overlevingskans van de mot. Dit kan de boomkleur zijn, maar minder direct duidelijke aspecten, zoals de vraag of de predator van de mot visueel jaagt, spelen ook een rol.

Wat ik probeer duidelijk te maken, is dat de waarde van eigenschappen van een levend wezen slechts een waarde is bij de gratie van de toestand van de omgeving. De waarde van een eigenschap is dus veranderlijk in zowel de tijd als in de ruimte. Dit raakt direct aan de grootste en meest algemene misverstanden over evolutie, namelijk dat deze een progressief verloop en een doel heeft. Met progressief bedoel ik de veronderstelling dat evolutie een

vaststaand verloop heeft, van primitief of eenvoudig naar steeds grotere complexiteit. Evolutie heeft echter geen doel. Evenmin kan volgens mij het concept van primitiviteit in deze context gebruikt worden, zoals bijvoorbeeld met de zeer populaire term “levend fossiel” wordt gedaan om soorten aan te duiden die al heel lang bestaan. De volgens deze opvatting meest primitieve organismen, de bacteriën, behoren tot de meest succesvolle levensvormen op aarde, gemeten naar diversiteit en geografische verspreiding. Een organisme dat in het heden leeft is -per definitie- modern en zeer goed aangepast. Op aarde heeft het leven geen kans om achterop te raken. Wie dat overkomt wordt direct verzwolgen in de stormen van de *struggle for life* en *survival of the fittest*.

XI.5. Het onderwerp van dit proefschrift: Genomica

Vrijwel iedereen heeft tegenwoordig wel eens van het menselijke genoomsequentie project gehoord maar niet veel mensen weten wat dit eigenlijk inhoudt. Met genoom wordt bedoeld de totale hoeveelheid DNA van een organisme. Het doel van een genoomsequentie project is het bepalen van de basenvolgorde van alle DNA van een organisme. Deze basenvolgorde-bepaling, sequencing in het Engels, is slechts de aanloopfase voor het werk dat binnen het onderzoeksveld van de genomica valt. Genomica kan omschreven worden als de studie naar de structuur, functie, en evolutie van genomen. Ik zal op de verschillende aspecten van deze definitie wat nader ingaan, dit binnen de context van dit proefschrift.

De studie naar de structuur van een genoom houdt ten eerste de bepaling van de basenvolgorde van al het DNA van een organisme in. De hoeveelheid DNA verschilt per organisme, evenals het aantal chromosomen; bij de mens is al het DNA, met een totaal van ongeveer 3 miljard basenparen, verdeeld over een 23-tal chromosomen, terwijl *Bacillus subtilis*, de bacterie waar ik onderzoek aan heb gedaan, slechts één chromosoom heeft, bestaande uit ruim vier miljoen basenparen. In het kader van het *B. subtilis* genoomsequentie project is in onze groep de DNA-volgorde bepaald van een 171,812 basenparen tellend fragment van het chromosoom.

Na het bepalen van de DNA volgorde is het eerste wat een onderzoeker meestal doet het identificeren van mogelijke genen. Een potentieel gen wordt geïdentificeerd door te zoeken naar doorlopende leesframes, series van codons van, bijvoorbeeld, ten minste 50 die niet onderbroken worden door een stopcodon (zie ook Figuur XI.2). In het algemeen ligt de grootte van een gen tussen de 50 to 1000 codons, maar kleiner of groter kan ook. Op het chromosoom van *B. subtilis*, dat ongeveer 4 miljoen basenparen bevat, zijn op deze manier ongeveer 4100 genen geïdentificeerd, terwijl naar schatting het aantal genen van de mens op ±100.000 zal uitkomen. Het is niet zo dat de mens, naar de hoeveelheid DNA, evenredig veel genen heeft ten opzichte van *B. subtilis*: in het *B. subtilis* chromosoom wordt er gemiddeld per 1000 basenparen 1 gen gespecificeerd, terwijl bij de mens dat getal ongeveer 1 gen per 30.000 basenparen is. Bij mens - en dit geldt voor alle eukaryoten - is de erfelijke informatie veel minder efficiënt opgeslagen dan bij prokaryoten (bacteriën). Genen van eukaryote organismen zoals de mens worden, in tegenstelling tot bacteriële genen, veelvuldig onderbroken door DNA-volgordes die niet voor de aminozuur-volgorde coderen; introns worden deze volgordes genoemd. Ik zal hier echter niet verder op in gaan. In het gebied

waarvan wij de basenvolgorde hebben bepaald zijn op de bovenstaande manier 170 potentiële genen geïdentificeerd.

Na het identificeren van de mogelijke genen door middel van het zoeken naar leesframes, worden de door die genen gespecificeerde aminozuur-volgordes vergeleken met alle bekende aminozuur-volgordes in publieke databanken. Gelijkenissen (homologieën) tussen aminozuur-volgordes duiden meestal op evolutionaire verwantschappen. Dit wil zeggen dat de genen die de homologe eiwitten specificeren een gemeenschappelijke voorouder hebben gehad. Twee genen die voor homologe eiwitten specificeren hoeven echter niet *per se* in twee verschillende soorten aanwezig te zijn. Eén van de belangrijkste lessen die tot nu toe uit genoomprojecten getrokken kan worden, is dat in het genoom van één soort altijd vele onderling homologe eiwitten gespecificeerd zijn. Dit kunnen zelfs groepen van tientallen eiwitten zijn die alle onderling homoloog zijn. Twee definities zijn geïntroduceerd om de verschillende typen homologiën van elkaar te onderscheiden. *Orthologen* zijn homologe eiwitten met een gemeenschappelijke voorouder die door speciatie (soortsvorming) gescheiden zijn geraakt. *Paralogen* zijn homologe eiwitten die ontstaan zijn door gen-duplicaties binnen een soort. Dit is in Figuur XI.6 geïllustreerd. In hoofdstuk IX staat het onderzoek beschreven aan vier paralogen die alle een vergelijkbare, maar net niet identieke functie vervullen in het eiwitsecretie proces.

Uit homologie-onderzoek is veel informatie te halen. Ten eerste kan, als een homologe aminozuur-volgorde wordt gevonden met bekende functie, een idee verkregen worden over de mogelijke functie van het onbekende eiwit. Ten tweede kunnen, als meerdere homologe eiwitten gevonden worden, mogelijke essentiële aminozuren en functionele eiwit-eenheden (domeinen) in de aminozuur-volgorde geïdentificeerd worden. Verder kunnen met dit soort analyses niet alleen inzichten verkregen worden in de evolutionaire geschiedenis van genen afzonderlijk, maar ook in de evolutie van soorten en zelfs groepen van soorten.

Naast het zoeken naar homologe eiwitten in hetzelfde of in andere organismen is het ook mogelijk op een andere manier aminozuur-volgordes te analyseren. Het is bijvoorbeeld mogelijk om aan de hand van bepaalde componenten in de aminozuur-volgorde van een eiwit te voorspellen of het eiwit gelocaliseerd zal zijn binnen de cel (cytoplasma), of in de celmembraan, of dat het in het medium zal worden uitgescheiden.

Als tenslotte alle genen zijn geïdentificeerd en hun verwantschappen en functionele componenten geanalyseerd, kan in het laboratorium verder gegaan worden met de zoektocht naar de functie van de genen. Als door de computer-analyses een idee is verkregen van de mogelijke functie van een gen kan dit natuurlijk heel gericht worden onderzocht. Hoofdstuk zes van dit proefschrift is hier een typisch voorbeeld van. Onder andere op grond van de homologie van het *yhxB* gen-product met bekende eiwitvolgordes kon afgeleid worden dat het YhxB-eiwit waarschijnlijk betrokken is bij de synthese van een celwand-component, teichoïnezuur. Deze aanname is vervolgens bevestigd via analyse van een bacteriestam waarin het betreffende gen was geïnactiveerd. Bij genen waar nog geen mogelijke functie voor is gevonden via de computer-analyses, gaat de zoektocht naar een functie volgens een vaststaand protocol. Het te onderzoeken gen wordt geïnactiveerd (uitgeschakeld). Vervolgens wordt onderzocht of het organisme met het mutante gen geschaad is met betrekking tot een

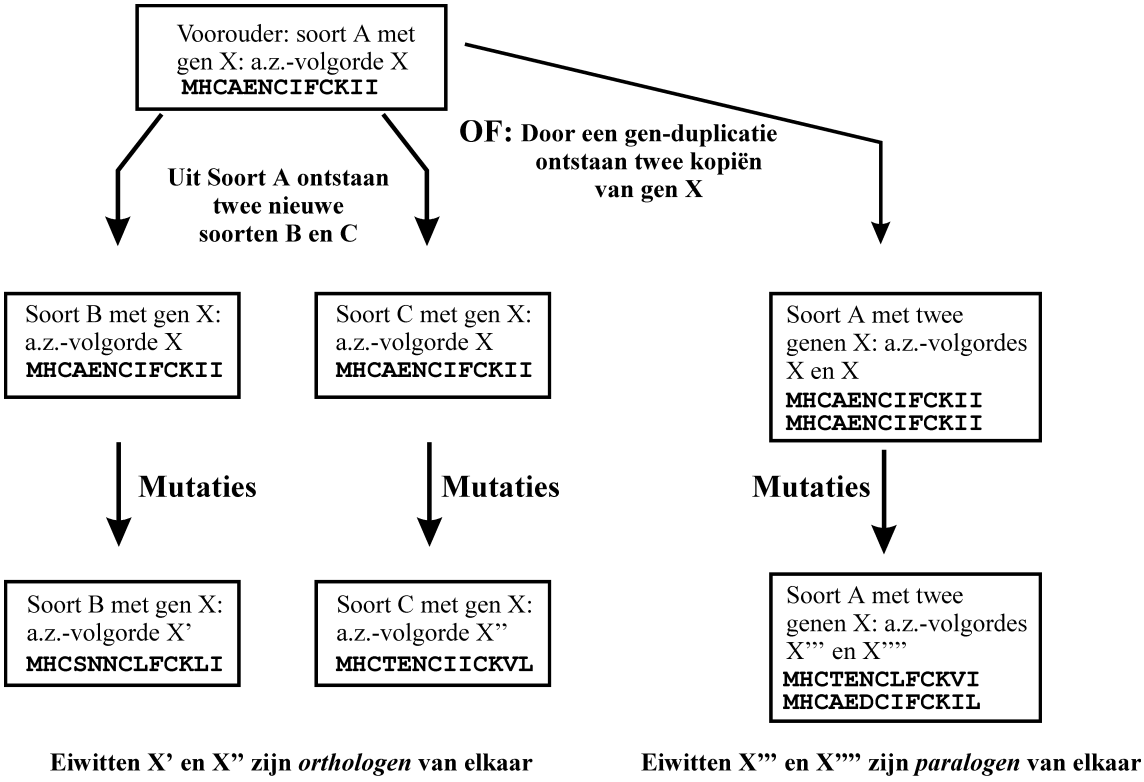


Fig. XI.6. Homologe eiwitten kunnen op twee manieren ontstaan. Als gedurende de evolutie twee nieuwe soorten ontstaan uit een gemeenschappelijke voorouder, zullen door mutaties de aminozuur-volgordes van de eiwitten in de nieuw ontstane soorten langzaam ten opzichte van elkaar veranderen (links in de figuur). Als dit type homologie wordt gevonden, tussen eiwitten uit verschillende organismen, spreken we van *orthologen*. Als binnen het genoom van één soort een gen-duplicatie (verdubbeling) plaatsvindt, zullen deze ook door mutaties langzamerhand van elkaar gaan verschillen. Dit type homologe eiwitten, die door twee verschillende genen van één organisme worden gespecificeerd, worden *paralogen* genoemd. a.z. = aminozuur.

aantal centrale processen, zoals de mogelijkheid om te groeien onder bepaalde omstandigheden of met bepaalde voedingsstoffen. Vervolgens poogt men de mogelijke functie(s) van het gen in te perken totdat de precieze functie is gevonden. Hoofdstuk zeven is een voorbeeld van deze werkwijze. Het zal duidelijk zijn dat dit gedeelte van het genomica-onderzoek verreweg de meeste tijd vergt.

Table of abbreviations

a.a.	amino acid
Ap^r	Ampicillin resistance
bp(s)	basepair(s)
CAT	chloramphenicol acetyltransferase
Cm^r	Chloramphenicol resistance
d.b. (D.B.)	DataBase
DNA	deoxyribonucleic acid
DTT	dithiothreitol
Em^r	Erythromycin resistance
IPTG	isopropyl ß-D-thiogalactopyranoside
kb	kilobase
kD (kDa)	kilodalton
Km^r	Kanamycin resistance
MM	Minimal Medium (as defined by *B. subtilis* Functional Analysis program)
M.U.	Miller Unit; definition of ß-galactosidase activity: (nmol ONPG)*$(min)^{-1}$*$(mg\ protein)^{-1}$
mtDNA	mitochondrial DNA
NB	Nutrient Broth (as defined by *B. subtilis* Functional Analysis program)
nt	Nucleotide
OD_{600} (OD)	Optical Density at 600 nm
ONPG	o-nitrophenyl ß-D-galactopyranoside
ORF	open reading frame
PAA	polyacrylamide
PAGE	polyacrylamide gel electrophoresis
PCR	Polymerase Chain Reaction
PMSF	Phenylmethylsulfonyl fluoride
RBS	ribosomal binding site (Shine-Dalgarno sequence)
rDNA	ribosomal DNA
RNA	ribonucleic acid
SASP	Small acid-soluble spore protein(s)
SDS	Sodium dodecyl sulfate
Sp^r	Spectinomycin resistance
w.t. (W.T.)	wild-type
X-gal	5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside

STELLINGEN

behorende bij het proefschrift:

Genomics in *Bacillus subtilis*

van Michiel Noback

1. Brown *et al.* (1990) zijn te prematuur geweest met hun aanname dat de gevonden carboxy-terminale afwijkingen in aminozuurfrequenties van een set *Escherichia coli* genprodukten effecten reflecteren die van toepassing zijn op alle prokaryoten. Brown *et al*. (1990). *Nucleic Acids Research* 18, 2079-2085.

2. Het genoom van *Mycoplasma genitalium* is, in tegenstelling tot wat Fraser *et al.* (1995) beweren, niet “The minimal gene complement”. Fraser *et al.* (1995). *Science* 270, 397-403.

3. Eiwitten die sterk geconserveerd zijn in vele organismen, zijn, in tegenstelling tot de aanname van Mushegian & Koonin, niet per definitie essentieel, zoals is aangetoond met het universele *hit* gen in dit proefschrift. Mushegian & Koonin (1996). *Proc. Natl. Acad. Sci. USA* 93, 10268-10273.

4. Myers’ bewering over de door de mensheid veroorzaakte massale uitsterving, die momenteel de biodiversiteit van de wereld reduceert, door hem als volgt verwoord: “The new results indirectly throw light on an overlooked but significant angle of the biotic ***crisis***: its grossly ***disruptive*** impact on the future course of evolution” bevat twee waarde-oordelen die -helaas- slechts betrekking hebben op de mens zelf. Massale uitstervingen, wat de oorzaken ook mogen zijn, zijn niet ontwrichtend voor de evolutie, maar eerder zeer stimulerend. Meyers (1997). *Science* 278, 597-598.

5. De uitspraak van Stephen Hawking “Wanneer in het heelal alles op een fundamentele manier van al het andere afhangt, dan is het wellicht onmogelijk om door onderzoek van geïsoleerde onderdelen van het probleem tot een volledige oplossing te komen” is, hoewel wij dat uit praktische overwegingen liever vergeten, zeker ook op de biologisch onderzoek van toepassing. Hawking (1998). Uitgeverij Bert Bakker, Amsterdam, Nederland.

6. De mens heeft in zekere zin minder wijsheid dan een bacterie; de laatste heeft een plan klaarliggen voor de slechte tijden die (altijd) komen.

7. Zo scherp als de arend te kunnen zien is slechts dán nuttig, wanneer men weet wat men zoekt.

8. Het probleem met tot de verbeelding sprekende wetenschappelijke concepten, zoals evolutie als progressief proces, is dat ze, lang nadat ze wetenschappelijk ontkracht zijn, nog als vaststaand feit bij het grote publiek voortleven.

9. Het geblinddoekt verzamelen van grote hoeveelheden gegevens kan toch lonend zijn, uiteindelijk.
   Dit proefschrift

10. In tegenstelling tot *in vivo* en *in vitro* experimenten, wordt de waarde van *in silico* en *in cerebro* experimenten zwaar onderschat.

11. Het is zeer waarschijnlijk dat inslagen van grote meteorieten op aarde de belangrijkste effectoren zijn geweest voor de loop van de evolutie.

12. Dromen zijn geenszins bedrog.

13. AATGCT GGTGAAGATGCTAATGAA GCTCGTTAGGAAATTGAT ATTAGT CATGAAACT GGTTTGGAAGAT CGTTTGAGTACTGAAAAT